

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

END OF THE BEGINNING

Final phase of 1000 Genomes
Project maps human genetic
variation in open-access
resource **PAGES 52, 68 & 75**

REVIEWS

AUTUMN BOOKS SPECIAL

*Fizz wars, geopoetry,
mind and matter*

PAGE 34

MAUNA KEA OBSERVATORY

MOUNTAIN DIFFICULTIES

*Is the Thirty Meter Telescope
a step too far?*

PAGE 24

THE HUMAN GENOME

25 YEARS OF BIG BIOLOGY

*Three major players reflect
on lessons learned*

PAGE 29

NATURE.COM/NATURE

1 October 2015 £10

Vol. 526, No. 7571



9 770026 063097

THIS WEEK

EDITORIALS

GOALS The UN sustainable development targets need research back-up **p.6**

WORLD VIEW Science careers must change to exploit impact **p.7**



FOOD FOR THOUGHT How different diets affect gut health **p.9**

Testing times

The unfolding Volkswagen saga highlights the need for better funding of regulatory science — and should prompt regulators to keep a closer eye on whether their rules are working.

Among the questions raised by the scandal that allowed the German car maker Volkswagen to sell 11 million vehicles containing software that cheats emissions tests, many will ask why the regulators failed to notice and halt the practice. The answer is not complicated. Regulated industries exert massive, discreet pressure on regulators such as the US Environmental Protection Agency (EPA), to stop them doing their jobs properly.

The research community has an opportunity here. It must use the Volkswagen crisis to highlight a broader problem: how regulatory science is funded, conducted and used. Long a poor relation of more prestigious investigations, this brand of applied science plays a crucial but much-neglected part in enforcing rules and saving lives.

It was a small academic team led by Daniel Carder, an engineer at West Virginia University's Center for Alternative Fuels, Engines and Emissions at Morgantown, that did the real-world 2012 emissions tests which brought the Volkswagen case to light. The work was paid for by a small grant from the International Council on Clean Transportation (ICCT) in Washington DC, a non-profit outfit of the type that many in the scientific and political establishments are inclined to disdain.

The ICCT was set up in 2001 "as a counterweight to the influence of the global automobile and energy industries in policy debates" and is staffed by several former employees of the EPA, the regulator responsible for policing car emissions in the United States. The EPA has a research and development budget of US\$537 million this year. The US National Institute of Environmental Health Sciences, part of the National Institutes of Health, has a budget of \$665 million. The budget of the European Union's Joint Research Centre — which, to be fair, had already published work highlighting flaws in emissions-testing regimes — is about €330 million (US\$371 million).

Why, then, does it take a \$50,000 grant from an obscure non-profit organization to expose what seems to be a systematic and widespread effort by Volkswagen, going back at least to 2009?

Almost every public discussion about industry regulation and the regulatory science that supports it concerns 'regulatory reform': a euphemism, in far too many cases, for the relentless process whereby those who are regulated push back against the regulator.

With exquisite timing, for example, Jeb Bush, the former governor of Florida and possible Republican nominee for next year's US presidential election, published an opinion piece on 22 September — perhaps written before the Volkswagen scandal broke — promising to regulate the regulators. He singled out EPA rules on clean water and carbon dioxide for repeal. "We are a nation of free men and women who are capable of achieving far more than liberals and regulators believe possible," Bush grandly declared.

It would be wrong, however, to suggest that only conservatives such as Bush encourage regulators to be bullied. Everyone has been at it. In Europe, for example, successive governments in France, the United Kingdom and Germany have each been lobbying the European

Commission for years, to block the planned introduction of more-realistic emissions tests for diesel engines.

Since the findings went public, it has emerged that the EU Joint Research Centre had already conducted tests that produced damning indictments of the existing regulations — if not of the vehicle companies. The EU is now moving ponderously towards more rigorous, on-road testing of car emissions, due to be introduced in 2016.

"The EPA is not ensuring the efficacy of its own regulations. That can and should change."

Who is best placed to conduct important regulatory science? It is not going to be done by the regulated industries or by academics who want to pursue friendly relations with those industries. (One positive side effect of the scandal could be to highlight the extent to which even companies with good public reputations, such as Volkswagen, carry agendas.)

Work that second-guesses the regulators is also unlikely to be supported by 'pure' science agencies, such as the US National Science Foundation. These agencies tend to avoid regulatory science because it is politically risky, as well as being prone to dismissal by programme managers as routine, or uninteresting.

There are two possible solutions. Basic-research agencies could open up more funding calls devoted expressly to regulatory science. Most politicians would resist that, but given recent events, some might support it. And regulators themselves need to ask tougher questions about how their rules are being implemented. The serendipitous nature of the Volkswagen case — in which the problem was brought to the attention of California and federal regulators by the Carder team's investigation — suggests that, for whatever reason, the EPA is not ensuring the efficacy of its own regulations. That can and should change.

This unfolding saga should, at least, lend regulators more heft and political support in the never-ending battle with their crafty and well-resourced charges. ■

Variety of life

An effort to sequence thousands of people's genomes reaches the end of the beginning.

"Nature is an endless combination and repetition of very few laws," said the nineteenth-century US poet Ralph Waldo Emerson. "She hums the old well-known air through innumerable variations."

Modern science has a good grip on most of those very few laws that drive life forward, most tellingly on how genetic material copies itself

from parent to offspring. The innumerable variations however? Not so much. They are, after all, innumerable.

That does not mean that science is not trying, and on pages 68 and 75 of this issue, *Nature* publishes the latest progress reports from this colossal effort. The papers mark the completion of the 1000 Genomes Project, the largest work yet to sequence the genetic information of hundreds of individuals in an attempt to tune into Mother Nature's hum of human variation. It completes a set of genomic reference tools — resources of genetic data produced by international collaborations — that dates back 25 years to the start of the Human Genome Project.

The bigger job, of tracking the relationships between genetic variation and human disease to help to develop effective treatments, is not finished, and may never be. But it is important from time to time to acknowledge and celebrate landmarks of achievement along the way. This week marks one such landmark.

The data sets produced by the 1000 Genomes Project are already in use. The genetic details of the volunteers provide a publicly owned and openly available asset in the era of big data, and offer a foundation for further study. Applications range from hunts for the genetic roots of human illness to analyses of population genetics and evolutionary history.

As technology continues to improve, so does the ability to capture genetic variation worldwide. The research published this week demonstrates that neatly. For a start, the eponymous 1,000 genomes analysed have extended to more than 2,500. The data now come from 2,504 individuals, across 26 distinct populations. From Chinese immigrants in downtown Denver, Colorado, and the Luhya tribe in Kenya to Punjabis in the dusty streets of Lahore, Pakistan, much of human life and diversity is here. The genetic data have been analysed more thoroughly than was possible before, which throws more light on rarer forms of variation.

The take home message: although most common genetic variants are shared across populations, rarer variants are often restricted to closely related groups. Many more rare variants are still to be identified.

The improved precision provided in this latest data set has also

enabled a more comprehensive map of structural variation across the human genome. For the first time, this includes analysis of eight structural-variation classes.

What now? Sequencing projects should continue to cast the net wide, and extend it further, to seek volunteers from regional and ethnic groups that are currently under-represented in global genetic databases. Meanwhile, the astonishing increase in genetic sequencing ability — even when compared with when the 1000 Genomes Project began in 2007 — has shifted the research bottleneck from generation of data to analysis and interpretation. Two challenges are to make sense of the non-coding regions of DNA and to tease out the links between genetic variation and clinical symptoms.

To exploit the gathered genetic information, more projects need to link and cross-reference it to clinical information and well-characterized phenotype data sets. On page 82, the UK10K Consortium publishes an early example of the latter: the first large-scale demonstration of whole-genome sequencing linked to complex traits.

As links to health records are established — and some, such as the UK Biobank study and the US Precision Medicine Initiative, are already on the books — it is crucial that public trust is secured. The ways in which scientists collect, store and share sensitive personal information must continue to evolve to ensure adequate safeguards. The Global Alliance for Genomics and Health has offered promising alternatives and a model to follow.

The final goal remains to make this flood of population-level genetic research relevant to personal health. Emerson would have approved. He was a proponent of individualism, a political philosophy that emphasizes the moral worth of the individual. He celebrated the non-conformist. And when it comes to the few laws that dictate the repetition of genetics, it is not just the 2,504 people whose variation is detailed this week who are the non-conformists. We all are. ■

Goals galore

The latest global targets from the United Nations must be translated into realistic policies.

The nations of the world approved a new development agenda in New York over the past weekend. The United Nations' 17 Sustainable Development Goals cover topics ranging from poverty reduction to environmental sustainability, and are accompanied by 169 detailed targets that are intended to help governments and aid organizations to focus resources. It is a noble initiative, in principle, and the world would undoubtedly be a better place by its target year of 2030 if these goals were met. But despite the promotional efforts — one of the main side events over the weekend culminated in pop diva Beyoncé and the rock band Pearl Jam performing Bob Marley's 'Redemption song' in New York's Central Park — it remains unclear what impact the goals will have on global affairs.

One problem is that there is a sense of déjà vu here. Back in 1992, the world set out a 351-page manifesto for human justice and environmental sustainability at the Earth Summit in Rio de Janeiro, Brazil. Eight years later, the UN adopted the eight Millennium Development Goals, which included halving extreme poverty rates and achieving universal primary education by 2015.

The aspirational agenda is still clear, but so too are the barriers to investment — they include corruption, political instability, poor

education systems, malfunctioning regulatory systems and the lack of a skilled workforce.

The real challenge is to identify and implement realistic policies that will get us where we say we want to be, and this is where academics must engage. The next step for the Sustainable Development Goals is to identify a range of health, economic and environmental indicators that can be used to track progress.

That debate is expected to extend into next year, and researchers should work to ensure that governments are collecting and reporting data. Scientists and policymakers must also redouble efforts to identify effective — and politically viable — strategies in which to invest a limited supply of money. Increasingly, development economists are doing just that, complete with rigorous testing, but there is scope for much more research in this important field.

The first and perhaps biggest opportunity to address some of these issues in a significant way will come when global leaders converge on Paris for the UN climate summit this December. Attempts to develop a telling international climate regime have languished for a quarter of a century, but there are signs of life, and governments around the world — rich and poor alike — are beginning to engage. The world is unlikely to see a single solution emerge, but the summit could produce a framework that will push all governments to invest in the policies, as well as in the science and technology.

Trillions of dollars of investment over the coming decades, public and private, are on the table. Directing that money to the right technologies and the right places would go a long way towards improving lives. ■

➔ **NATURE.COM**
To comment online,
click on Editorials at:
go.nature.com/xhunqv



Science must prepare for impact

To maintain public support, researchers need to be able to adapt to the rapidly changing needs of society and politicians, warns Guy Poppy.

What do scientists do for society? Some researchers may resent the increasing calls for them to demonstrate ‘impact’, but my time seconded to the UK Food Standards Agency as chief scientific adviser has convinced me that such pressure will only increase. Policymakers are no longer willing to hand over billions of pounds of taxpayers’ money to scientists in exchange for a vague promise that something good will come from it.

What politicians and society expect from science is changing rapidly, and science must change with it, or risk losing public support. Academics and leaders of the scientific community must realize that the system is failing to prepare researchers to meet wider society’s requirements.

At present, the metrics of scientific success used by most universities — citations, publications and grant money — encourage a linear career path from postgraduate studies to a tenured position. The bottleneck in this process — the oversupply of PhD students that cannot be found full-time jobs in academia — has received much attention. (Although in many ways it is good for society that most of these people do not pursue academic careers.) But we need more awareness of a second problem, which affects those who do continue into academic jobs. This linear track creates successive generations of scientists who are unable or unwilling to demonstrate the kind of societal impact that policymakers demand, industry requires or society needs.

When pressed to provide examples of impact, most senior researchers can do so. The latest Research Excellence Framework assessment of universities in the United Kingdom, for example, provided hundreds of case studies, including of start-up companies and of policy that has informed research. But as the demands for impact increase, we need more early-career scientists who are able to spend time working across academia, industry and policy.

Under the present system, researchers who frequently swap between academia, industry and policy are rare. They tend to have pursued such careers by chance and determination, rather than by design. That is my story: although my career was advancing on many fronts — agro-ecological research, industrial collaboration and policy relevance — my performance was typically measured using the standard academic metrics.

I made it, but I fear others may not. This route is not attractive to many, so it limits the supply of people into this important and growing job profile, which in turn is not good for society.

Already a super-competitive career path, it will become even harder for young scientists to dare to step from the tenure track — for a secondment to a policy think tank for example. Although natural selection may allow some individuals to succeed in this rapidly evolving research environment, it might

narrow the ‘phenotype’ and so reduce the diversity of people and styles, which is important for a resilient workforce.

A similar dynamic holds back interdisciplinary research. Many young academics are excited by the idea but fearful of moving from a discipline-based assessment and reward system. Interestingly, the rapidly growing demand for interdisciplinary research is largely in response to the need to address the grand challenges facing society or to deliver a knowledge-based economy.

How can nonlinear career tracks be encouraged? PhD training is leading the way. There are industrial PhDs, and professional internships are now common in doctoral training partnerships.

But once a scientist starts off in an academic career, the incentives dwindle. More universities need to develop metrics to assess and

reward performance in areas other than grants and papers. Increased tuition fees for students in the United Kingdom have already focused attention on teaching ability. Case studies of impact should also be measured and rewarded. Some land-grant universities in the United States already measure how their agricultural researchers contribute to the state economy.

Industry and policymakers should welcome academics’ attempts to engage in applied research that has impact. And academics must drop the snobbery that holds back work on such topics.

Frequent exchange of people and ideas will show academics how to ensure that their research has impact, and will allow those not in universities to access and use the knowledge created in them. And as such interactions increase, those

who lecture students and prepare them for the workplace will start to obtain a better understanding of what is required.

It is not just the scientific community that can change to encourage a new breed of scientists who are comfortable with regular transitions between job types. I know from experience that it is often hard to achieve progression and promotion within the UK scientific civil service. Perhaps schemes that encourage more-permeable walls between academia, industry and policy could help to recruit and retain scientists in government roles, too. Together with colleagues, I am working on improving links between the civil service and universities.

We have made progress with interdisciplinary research, although not enough. Postgraduates have more opportunities, and it is now time to improve the career options for academics. Science is complex, and so are knowledge and solutions. It is no longer realistic — or sustainable — to insist that scientific careers be simple to manage and judge. ■

Guy Poppy is professor of ecology at the University of Southampton, UK, and chief scientific adviser to the Food Standards Agency.
e-mail: g.m.poppy@soton.ac.uk

**RESEARCHERS
WHO FREQUENTLY
SWAP
BETWEEN ACADEMIA,
INDUSTRY AND
POLICY ARE
RARE.**

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/6dhjee

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

NANOMATERIALS

Sunblock stays on skin surface

Using nanoparticles to encapsulate the ultraviolet (UV) filters found in sunscreen might prevent them from being absorbed by the skin — and could even improve their UV-blocking performance.

Some studies have shown that chemical UV filters have negative effects on cells when they penetrate skin. To stop this absorption, Mark Saltzman and his colleagues at Yale University in New Haven, Connecticut, coated a typical UV filter — padimate O — with nanoparticles that have sticky aldehyde groups on their surfaces. The coated UV filters stuck to the skin of mice and pigs even when exposed to water, and the nanoparticles prevented the filters from penetrating the skin.

Sunblock that used these nanoparticles and contained only 5% of the amount of UV filters found in conventional sunblock absorbed the same level of UV radiation.

Nature Mater. <http://dx.doi.org/10.1038/nmat4422> (2015)

ANIMAL BEHAVIOUR

Fish launches jaw to feed on land

A species of fish has an unusual way of eating — it thrusts its jaw out and downwards to nab prey on land.

Krijn Michel at the University of Antwerp in Belgium and his colleagues took high-speed video and made 3D reconstructions of the largescale four-eyed fish (*Anableps anableps*), which feeds from mudbanks. They found that the fish extends and rotates its upper jaw towards the ground while it turns its

lower jaw downwards at a right angle, allowing it to clamp its mouth around its prey.

This mechanism differs from those of other land-feeding fish, which either curl their whole bodies downwards or pivot on their fins towards prey. *J. Exp. Biol.* 218, 2951–2960 (2015)

CLIMATE CHANGE

Clean air puts Arctic ice in peril

Cleaner air in the high north could reduce Arctic sea ice by an area of about one million square kilometres this century.

Air pollution has a net

cooling effect on the climate, and has partially offset the decline of Arctic sea ice since the mid-1970s. John Fyfe and his colleagues at the Canadian Centre for Climate Modelling and Analysis in Victoria, Canada, used an Earth-system model to simulate sea-ice changes in the twenty-first century with and without projected reductions of global aerosol emissions. Cleaner air accounted for 15–40% of the Arctic ice melting simulated under a range of greenhouse-gas emission scenarios.

In a model with high greenhouse-gas emissions and large projected reductions in

of warmer temperatures and drier soils. The researchers measured the tongues of 170 bees and found that they have got shorter by an average of about two millimetres since the 1970s in two dominant bee species in that area, *Bombus balteatus* and *Bombus sylvicola*.

Shorter tongues allow bees to feed on nectar from a greater variety of flowers, rather than from just long-tubed blooms.

Science 349, 1541–1544 (2015)

PLANT BIOLOGY

CRISPR cripples plant viruses

Plants that have been engineered to contain the CRISPR–Cas9 system are resistant to viral infections that reduce crop yields.

The CRISPR system, first discovered in bacteria, uses



NICOLE MILLER-STRUTTMANN

EVOLUTION

Bee tongues shrink as climate warms

Bees in some parts of the US Rocky Mountains have evolved shorter tongues, probably in response to a decline in flower populations caused by climate change.

Nicole Miller-Struttman at SUNY College in Old Westbury, New York, and her co-workers studied bees at three alpine sites in the Rocky Mountains. Similar to other mountainous habitats around the world, the Rockies have seen a drop in the number of flowers because

certain RNA molecules as guides to recognize specific DNA sites in genomes that the Cas9 enzyme then cuts. Two groups of researchers have designed guide RNAs to target and disrupt DNA from geminiviruses, which infect many crops. Caixia Gao of the Chinese Academy of Sciences' Institute of Genetics and Developmental Biology in Beijing and her colleagues focused on the beet severe curly top virus. They found that transgenic CRISPR-Cas9 plants had 60–80% less viral DNA than control plants, and did not show disease symptoms such as leaf curling. Similarly, Daniel Voytas of the University of Minnesota in Minneapolis and his colleagues targeted the bean yellow dwarf virus genome and found 5–87% less viral protein in infected engineered plants.

This strategy could be used to develop disease-resistant transgenic plants, the teams say. *Nature Plants* <http://dx.doi.org/10.1038/nplants.2015.144> (2015); <http://dx.doi.org/10.1038/nplants.2015.145> (2015)

MICROBIOLOGY

Diet makes gut change speed

Interactions between diet and gut microbes affect how quickly food moves through the gut.

To simulate dietary changes that occur when people travel to places with different cuisines, Jeffrey Gordon at Washington University in St Louis, Missouri, and his team took germ-free mice and transplanted them with gut microbes from people consuming one of five different diets from around the world. They then fed the mice a series of all those diets and measured transit

times of dye-stained food through the gut. They found that transit time



varied with different combinations of diet and microbial community, and that it correlated with certain metabolites produced by some bacteria.

Turmeric (pictured), a common ingredient in Bangladeshi food, in particular decreased gut motility in mice carrying microbes and eating food from Bangladesh — in part by increasing the production of bile acid, which was converted by microbes into compounds that slow down gut movement. The approach could be used to identify components of different diets that affect gut health, the authors say. *Cell* 163, 95–107 (2015)

ECOLOGY

Creatures are busy in the polar night

The high Arctic is thought to be biologically quiescent during the long 'polar night' — the winter months when the Sun never rises. But Jørgen Berge at the Arctic University of Norway in Tromsø and his colleagues have discovered a surprising level of biological activity.

During three winters in the cold and dark in Kongsfjorden, Svalbard, the team recorded, for example, growing bivalves, foraging seabirds, scavenging crabs and reproducing and respiring zooplankton.

The ecosystem seems to thrive without photosynthesis by relying on energy that has been stored or brought in with Atlantic water.

Curr. Biol. <http://doi.org/7xd> (2015)

ROBOTICS

Robot moves when squished

Soft elastic materials that buckle in a vacuum can generate robot motions.

George Whitesides at Harvard University in

SOCIAL SELECTION

Popular topics on social media

Gender-disparity study faces attack

Bias against women in science is a well-studied and well-documented phenomenon. But some cases may not be as clear cut as they first seem. A study published this week in the *Proceedings of the National Academy of Sciences* claimed that female researchers in the Netherlands are more likely than men to lose out when applying for grants. The paper gained widespread support on social media, but some commenters quickly raised doubts. In a blog post, Casper Albers, a statistician at the University of Groningen in the Netherlands, argued that the authors had fallen victim to a common statistical error, which negates the main finding. But the paper's lead author Romy van der Lee, a psychologist at Leiden University in the Netherlands, says that she stands behind the team's conclusion that gender affects success. *Proc. Natl Acad. Sci. USA* <http://doi.org/7v9> (2015)

➔ **NATURE.COM**
For more on popular papers:
go.nature.com/i6zdza



Cambridge, Massachusetts, and his colleagues built soft actuators out of squishy cubes containing air pockets. They attached rigid components such as grippers or legs to the cubes and sucked the air out of the pockets using a vacuum. This caused the cubes to collapse, driving the motion of the attached robot parts. When the vacuum was removed, the cubes returned to their original shape. By repeatedly changing the applied pressure, the team made robots that could walk or grab objects (pictured).

The buckling actuators can also be stacked to allow for more-complex motions, the authors report. *Adv. Mater.* <http://doi.org/f3gcnp> (2015)

CANCER IMMUNOTHERAPY

Molecular switch controls therapy

A molecular 'remote control' could enable researchers to make a powerful cancer

therapy safer.

The therapy relies on engineered immune-system cells called T cells that recognize and kill tumours, and has shown promise in clinical trials. But the T cells can also attack and damage healthy cells. James Onuffer and Wendell Lim at the University of California, San Francisco, and their colleagues designed an approach in which the T-cell receptors that recognize cancer cells are split in two, and will only assemble and function when triggered by a compound similar to the drug rapamycin.

T cells engineered in this manner only attacked their target cells in mice when the compound was present. The approach could provide a way to modulate the timing and intensity of engineered T-cell responses in humans.

Science <http://doi.org/7v4> (2015)

➔ **NATURE.COM**
For the latest research published by Nature visit:
www.nature.com/latestresearch

SEVEN DAYS

The news in brief

PEOPLE

Plagiarism claims

Hanover Medical School in Germany is examining allegations that German defence minister Ursula von der Leyen committed plagiarism in her 1990 medical dissertation in obstetrics. The claims originated on 'VroniPlag Wiki', a website that scours academic theses for plagiarism. Two federal ministers have lost their PhD titles and posts following evidence of misconduct: Karl-Theodor zu Guttenberg, a defence minister, in 2011, and education and research minister Annette Schavan in 2013. The university says that it has launched a formal, preliminary investigation at the request of von der Leyen, who denies the scientific misdemeanour. See go.nature.com/dkpafr for more.

Greek governance

Laser physicist Costas Fotakis remains the minister for research and innovation in Greece's new government, after the left-wing Syriza party won a snap election. Fotakis, whose role was announced on 22 September, is expected to push forward his plans to turn around the country's moribund research landscape — goals that he has been developing since January, when he was first appointed to the position. The minister's promises include major calls for research proposals in October and, in the long term, the creation of Greece's first dedicated research fund.

EVENTS

Emissions scrutiny

Car makers in North America face increasingly stringent emissions testing in the wake of the news that Volkswagen fitted cars with software to

cheat tests. Additional checks for 'defeat devices' will mean that approvals for compliance with US emissions regulations will take longer, the US Environmental Protection Agency (EPA) announced on 25 September. The EPA will also work with Californian authorities and with Environment Canada to ensure that vehicles already on the road comply with emissions regulations under regular driving conditions. See go.nature.com/hz72x3 for more.

Polio-free Nigeria

The World Health Organization in Geneva, Switzerland, has officially

removed Nigeria from the list of countries where wild poliovirus still circulates. As of the 25 September announcement, Afghanistan and Pakistan are the only remaining countries that harbour the virus. In July, Nigeria celebrated a full year without a single new case of the disease.

Research at sea

The RV *Neil Armstrong*, a state-of-the-art research vessel, arrived at the Woods Hole Oceanographic Institution in Massachusetts on 23 September. The ship is a gift from the US Navy, part of the force's 70-year tradition of supporting

and black holes, during its five-year mission. ASTROSAT has four telescopes that will simultaneously study space in visible light, ultraviolet and low- and high-energy X-rays, as well as a sky-scanning monitor to detect transient X-ray emissions and γ -ray bursts. See go.nature.com/ago5tf for more.

marine studies by giving select institutions research vessels. The 73-metre ship, which can carry a 20-person crew and 24 scientists, is gearing up for its first science mission, planned for May 2016 in the North Atlantic. The *Neil Armstrong* replaces the RV *Knorr*, which has been used since 1970.

BUSINESS

Alaska drilling ends

The oil company Shell has halted its controversial offshore exploration in Alaska for the foreseeable future. The firm said on 28 September that mineral fuel reserves in the Burger J exploratory well in the



ARUN SANKAR/AP/PA

India's astronomy ambitions take flight

The Indian Space Research Organisation's first satellite dedicated to astronomy, ASTROSAT, launched on 28 September from the Sriharikota spaceport in the Bay of Bengal. With its five instruments, the observatory aims to study star-birth regions and high-energy processes, including binary star systems of neutron stars

NASA/JPL/UNIV. ARIZONA
Chukchi Sea are too limited and costs too high to warrant further oil and gas exploration in an “unpredictable federal regulatory environment”. The US approval in August of Shell’s drilling operations in the fragile Arctic region had caused an outcry among environmental groups that were concerned about a possible oil spill.

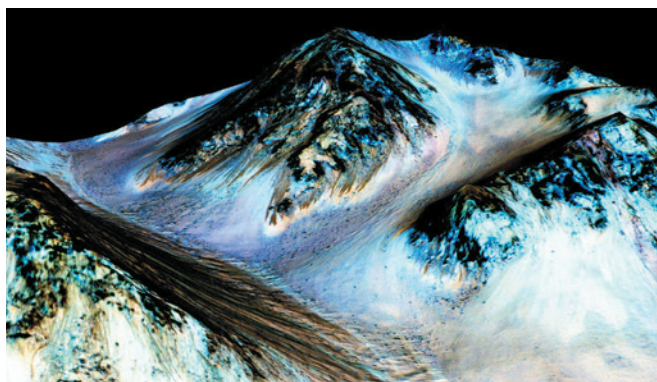
RESEARCH

Drug-study launch

The US National Institutes of Health has kicked off the largest-ever longitudinal study of adolescent drug use. On 25 September, the agency announced a US\$300-million award to 13 institutions for the Adolescent Brain Cognitive Development study, which will track substance use, mental health, brain structure and other factors in 10,000 children for 10 years (see *Nature* **512**, 123–124; 2014). The study, led by researchers at the University of California in San Diego, will track participants from age nine or ten, before they have started using drugs and through the period of highest risk for mental-health disorders.

Brine on Mars

Salty water exists today on Mars, NASA announced on 28 September. Data from the agency’s Mars Reconnaissance



Orbiter show that dark, shape-shifting streaks on some Martian slopes contain hydrated salts (L. Ojha *et al.* *Nature. Geosci.* <http://doi.org/7xw>; 2015). The streaks (pictured) have long been linked to the possible flow of water on the Martian surface, but the latest chemical analysis has produced the strongest evidence yet that there is liquid water on the red planet today — probably in damp, salty soil.

POLICY

China carbon plan

Chinese President Xi Jinping announced plans on 25 September to establish a national trading system to curb greenhouse-gas emissions, during a visit to US President Barack Obama in the White House. The cap-and-trade system, which sets a national ceiling on emissions and

allows companies to buy and sell emissions allowances as needed, will begin in 2017. The system is expected to cover emissions from the electricity sector and other energy-intensive industries, and will be larger than that of the European Union, which currently hosts the world’s largest carbon market. See page 13 for more.

Animal research

Transparency about UK animal research is improving, but several institutions are still not complying with a voluntary code of practice introduced last year, according to the London-based Understanding Animal Research (UAR) group. In its first annual report, published on 28 September, the group says that “clear progress” has been made since the May 2014 ‘concordat’ that commits signatories to openness. Of

COMING UP

4–8 OCTOBER

The International Society for Pediatric Neurosurgery meets for the 43rd time in Izmir, Turkey.
ispn2015.org

5–6 OCTOBER

Politicians and academics gather in Valparaíso, Chile, to discuss the future of the world’s oceans.
go.nature.com/rjxfdr

6–9 OCTOBER

The International Conference on Man–Machine Interactions convenes in the Beskid Mountains, Poland.
icmmi.polsl.pl

the 92 signatory universities and institutions reviewed in the report, 85 provided the UAR with details of their compliance. Although most institutes met or exceeded the required reporting standards on issues such as having public policy statements on animal use, several did not. UAR says that, by next year, non-compliance will result in expulsion from the concordat.

Emissions pledge

Brazil committed on 27 September to reduce its greenhouse-gas emissions to 37% below 2005 levels by 2025, aiming for a 43% reduction by 2030. Brazil is the first major developing country to commit to reducing absolute emissions, but environmentalists say that the pledge could have been stronger — the bulk of the reduction has already been achieved owing to a roughly 82% drop in Amazon deforestation since 2004. Brazil also promised to increase energy efficiency and expand renewable energy.

➔ **NATURE.COM**

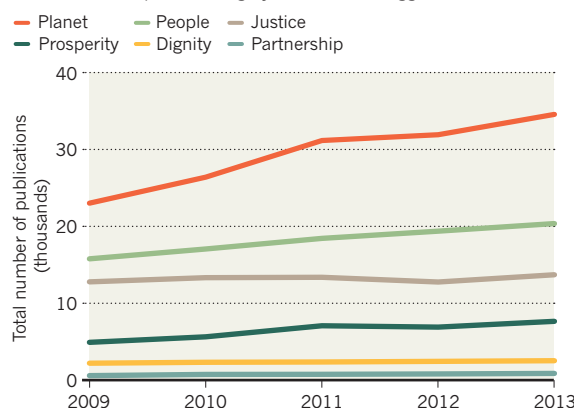
For daily news updates see:
www.nature.com/news

TREND WATCH

The field of ‘sustainability science’ is growing twice as fast as the average growth rate of research. A 24 September report by Elsevier and SciDev.Net used the Scopus database of research publications to identify trends based on six of the scientific themes that underpin the United Nations’ Sustainable Development Goals. Environment-related research (under the ‘planet’ theme) dominates overall, and ‘partnership’ research for catalysing global collaboration has seen limited growth.

SUSTAINABILITY RESEARCH MAKES PROGRESS

Publications on sustainability science have risen since 2009. Research in the ‘planet’ category has seen the biggest increase.



NEWS IN FOCUS

BRAZIL Science paralysed as economic crisis takes its toll **p.16**

GENE EDITING Chinese institute to sell micropigs as pets **p.18**

ANCIENT SCIENCE Model Universe recreated to test legend **p.19**



TECHNOLOGY Scientists seek a digital currency that betters Bitcoin **p.21**

KEVIN LAMARQUE/REUTERS



Chinese President Xi Jinping and US President Barack Obama have both made pledges on fuel efficiency.

CLIMATE CHANGE

China backs cap-and-trade

Climate commitment raises hopes for new global pact to limit greenhouse-gas emissions.

BY JEFF TOLLEFSON

With US President Barack Obama at his side, Chinese President Xi Jinping announced plans on 25 September for a national emissions-trading system to limit greenhouse-gas production.

The programme will begin in 2017. Xi's announcement builds on a 2014 climate agreement between the United States and China — the world's two largest emitters — and could help to build momentum for a new global climate pact ahead of United Nations talks in Paris in December.

"Our cooperation is delivering results for both our nations and the world," said Obama, speaking in the White House rose garden,

where he and Xi also announced a plan for reporting greenhouse-gas data. Experts say that this could make it easier to track energy use and emissions in China.

The two leaders also pledged to implement fuel-efficiency standards for heavy-duty vehicles by 2019, and to work together to promote global development.

Xi said that China will contribute 20 billion yuan (US\$3.1 billion) — an amount level with a comparable US commitment — to an international fund intended to help developing countries to address global warming. The announcement comes as countries gather at the UN headquarters in New York City to adopt new Sustainable Development Goals to guide poverty-reduction efforts until 2030.

This wave of US–China cooperation comes after similar commitments revealed by Xi and Obama at a meeting in Beijing in November 2014. Then, Xi said that China would increase its production of renewable energy and ensure that its greenhouse-gas emissions would peak around 2030.

Although questions remain about how high China's emissions could rise between now and then, observers say that the country's adoption of a market-based cap-and-trade approach to control emissions shows that it is serious about tackling global warming and air pollution.

"This is truly significant, and it's going to be extremely important in putting wind at the backs of delegates at the climate summit in Paris," says Robert Stavins, an economist ▶

► at Harvard University in Cambridge, Massachusetts.

Cap-and-trade systems set an overall limit on emissions and then distribute allowances to emitters. As total emissions decline, companies can buy permits to emit more — or make money by selling unused allowances. The goal of such a system is to spur innovation and to allow companies to find the cheapest way possible to reduce emissions.

The European Union has been operating the world's largest cap-and-trade system since 2005, and California and a group of northeastern US states run similar systems. China's trading scheme is expected to encompass roughly half of the country's emissions, including those from electricity generation and energy-intensive industries. That would make it the largest such system in the world.

GLOBAL CONNECTIONS

China's plans come as countries that will participate in the upcoming UN climate talks grapple with whether to allow links between emissions-trading systems in different parts of the world.

Stavins says that a coalition of countries, including Venezuela, Nicaragua and Cuba, is pushing to prevent international trading, which would prevent money from flowing to those regions where it is easiest — and cheapest — to reduce emissions. "That, in my mind, would be a tragedy," he adds.

Although many economists believe that cap-and-trade systems represent the cheapest way to reduce emissions, their track record in the real world is mixed. The EU carbon market, for instance, crashed early on owing to an oversupply of allowances, and even today prices remain around €8 (US\$9) per tonne of carbon dioxide — too low to spur a major revolution.

"This is truly significant, and it's going to be extremely important in putting wind at the backs of delegates at the climate summit in Paris."

China has been experimenting with carbon trading at the provincial and city level, but implementing a national system could be particularly difficult. Researchers have had a hard time verifying even basic information about energy consumption and emissions, and this could increase the risk of early mistakes, says Glen Peters, a climate-policy researcher at the Center for International Climate and Environmental Research in Oslo. He co-authored a paper published last month suggesting that Chinese emissions may be lower than previously believed (Z. Liu *et al.* *Nature* 524, 335–338; 2015).

"There will be mistakes, no doubt, but lots of learning by starting early," says Peters. "And when it is really needed to reduce emissions, a functioning framework will be in place." ■



Wildfires have proliferated during California's ongoing drought.

HYDROLOGY

California faces arid future

El Niño might bring relief, but droughts are likely to return.

BY ERIKA CHECK HAYDEN

Three brown, withered lawns surround David Behar's home in Marin County, north of San Francisco in California. Behar, who directs the climate programme at the San Francisco Public Utilities Commission, no longer waters his grass — after several years with next to no rainfall, he gave in and brought greenery to his home in the form of drought-resistant plants instead. It is just one of the many adjustments that Californians have had to make as the state enters its fifth year of drought (see 'Dry state').

As of 30 September — the end of the 2015 'water year' — the state's water supplies are desperately low. The spring snowpack is the paltriest ever measured — by April it contained just 5% of a normal year's

water — and by the end of August the major reservoirs held 59% of their historical average. Wildfires have burned through almost three times more land than they do in an average year. And there are myriad ecological impacts, including more patches of dead foliage than usual in the canopies of the state's iconic giant sequoia trees.

Yet despite the severe lack of rainfall, the state's biggest consumer of water has fared remarkably well. Agriculture last year generated revenues that were just 1.4% lower than in 2013, when it took in a record US\$34 billion, according to the Pacific Institute, a think tank in Oakland, California. Agriculture-related employment reached a record high of 417,000, and the amount of land being used for farming had fallen by less than 10% from pre-drought levels.

But that has come at a cost. Farmers have

storm pattern that has delivered extremely wet winters to California in the past. But plentiful rains are by no means certain, especially for regions outside southern California. "It remains to be seen whether an El Niño will provide relief this year," Behar says.

But the rains are unlikely to last for long. A team led by climatologist Noah Diffenbaugh of Stanford University in California has used historical data and climate models to show that global warming is increasing the odds of the state seeing warm, dry conditions similar to those that spawned the current drought (N. S. Diffenbaugh *et al. Proc. Natl Acad. Sci. USA* **112**, 3931–3936; 2015).

The droughts could even last for many decades. By incorporating palaeoclimate data into climate models, Benjamin Cook of the NASA Goddard Institute for Space Studies in New York City and two co-authors are predicting droughts that could last as long as 35 years (B. I. Cook *et al. Sci. Adv.* **1**, e1400082; 2015).

"We're in a new climate, and it's a climate in which the probability of severe drought conditions is elevated," Diffenbaugh says. "That recognition is really critical."

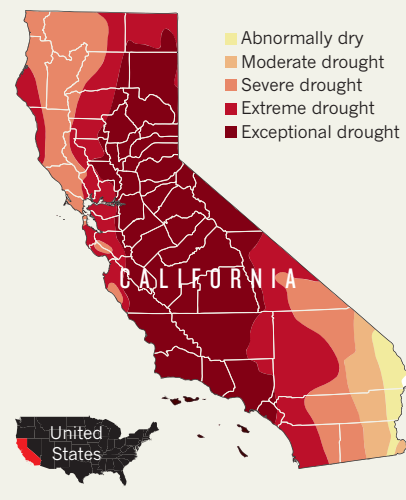
The state is working hard to respond to the dire warnings. In April, governor Jerry Brown called for a 25% reduction in municipal water consumption. Californians managed to save even more than that in both June and July, even though that is when irrigation needs tend to be highest. Last September, Brown signed legislation that takes the first steps towards regulating groundwater use by asking localities to make plans to ensure the sustainability of their groundwater supplies. However, some complain that he has not been tough enough. The path to sustainability does not need to be in place until 2040 — that is hardly an aggressive timeline, says Gleick.

And in June, the state's water board ordered some of the most-senior water-rights holders — the farmers and irrigation districts entitled to draw first in times of shortage — not to take water from rivers and streams. Some irrigators have responded with lawsuits that are still working their way through the courts.

The state's water-rights system is said to discourage conservation, because rights holders with priority may see their water allotments cut if they do not use their allocations in a given year.

DRY STATE

On 22 September, almost half of California was seeing exceptional drought conditions, and 25% was in extreme drought.



PUSH TO RECYCLE

Meanwhile, Behar and other planners around the state are pushing technological fixes that range from less-water-intensive agriculture and more-efficient home appliances to treating wastewater for reuse. Orange County already pumps treated wastewater into its groundwater; San Diego is developing a similar system and, on 8 September, the San Francisco Public Utilities Commission approved a plan to recycle wastewater to irrigate Golden Gate Park.

Steven Ritchie, the commission's assistant general manager for water, says that some regions are even looking into the long-discussed idea of desalination — a perennial option that is seldom used because it is energy intensive and therefore expensive. Ritchie says that one idea on the table for the region would be to cut the cost by desalinating brackish water from the San Francisco Bay Delta, rather than ocean water. Water produced in that way would cost about as much as that pumped from wells, Ritchie says, but would still be more costly than surface water. "The choice might be: expensive water, or no water," Ritchie says. "We have to continue to plan for the future and diversify our supply; we cannot take anything for granted." ■

sustained production mainly by pumping up huge amounts of groundwater. "The massive overdraft of groundwater to make up for lost surface water has buffered farmers, and that can't continue forever," says Pacific Institute president Peter Gleick. In August, NASA reported that the massive increase in pumping has caused parts of the state to sink by 33 centimetres in less than a year. Some households that rely on wells have been left without water to shower or wash dishes.

Efforts to regulate water use are being hampered by a lack of data on groundwater withdrawals, Gleick says. "We don't know who's using how much," he says. "There are really big gaps in the data."

CLOUDS ON THE HORIZON

The weather looks set to change, although only in the short term. The US National Oceanic and Atmospheric Administration predicts that this winter will see a strong El Niño, a

ZACKARY CANEPARI/NT/REDUX/EVINE



**MORE
ONLINE**

VIDEO OF THE WEEK



Bat has unique nectar-sucking trick
go.nature.com/w7mhue

MORE NEWS

- Academia.edu launches peer-review social network go.nature.com/cxku2i
- NASA images reveal Pluto's 'snakeskin' go.nature.com/pedbb9
- Protein promises to improve genome editing with CRISPR go.nature.com/d4gtir

NATURE PODCAST



Digital currency; antibiotics; and 25 years of cataloguing the human genome
nature.com/nature/podcast

NATURE VIDEO

AUSTERITY

Brazilian science paralysed by economic slump

From unpaid electricity bills to delayed participation in a telescope project, funding cuts bite.

BY ELIZABETH GIBNEY

For a decade, biochemist Octavio Franco has watched Brazil's economy boom and its research investment surge. "We're starting to do really high-quality science," he says. But as the nation's economy stalls, Franco is one of many who fear that this budding research ecosystem is in jeopardy.

In 2013, growth decelerated in what was once one of the fastest-growing economies in the world; in 2014, it all but stopped. Over the past year, the resulting cuts to federal and state science funding have paralysed research, says João Calixto, a biomedical scientist at the Federal University of Santa Catarina. "2015 has been a big mess," agrees Franco, whose two labs at the Catholic University of Brasília and the Catholic University of Dom Bosco in Campo Grande together employ almost 100 researchers. Although his labs have won 11 grants for 2015, they have received cash for just 2 of them.

The 2016 budget proposal, which Brazil's President Dilma Rousseff presented to Congress in September, only worsens the situation. It requests 24% less for the federal Ministry of Science, Technology and Innovation (MCTI) than did the 2015 proposal (see 'Federal funding woes'), and freezes new scholarships in a flagship exchange programme, Science Without Borders. The programme is close to meeting its goal of sending 101,000 Brazilian undergraduate and postgraduate students to top institutions abroad by the end of 2015, and aims to send a further 100,000 by 2018.

Federal science felt the squeeze in May, when Rousseff's administration chopped the MCTI's budget by almost 2 billion reais (US\$500 million), or some 25% (the ministry's spending limit has since shrunk further). Although the MCTI says that it is maintaining existing grants and fellowships, it dropped some regular calls for research proposals and is creating few new programmes. This year, there will be no 'universal call' from the National Council for Scientific and Technological Development (CNPq), a R\$200-million funding stream that last year was allocated to more than 5,000 projects. Junior postdoctoral fellowships part-funded by the CNPq seem to have been frozen from September too, says Franco.

Institutions are struggling to keep their labs going, says Cassio Leandro Barbosa, an



Proposed cuts to government spending have provoked demonstrations in Brasília.

astronomer formerly at the University of Paraíba Valley in São Paulo state. A 75% cut to a Ministry of Education graduate programme is making it tough to buy equipment, keep laboratories working and pay travel expenses for field research and meetings. "Some federal institutes don't have funds enough to pay the basics, like electric bills and cleaning," he says. The depreciation of the real has hiked the price of imports, including supplies,

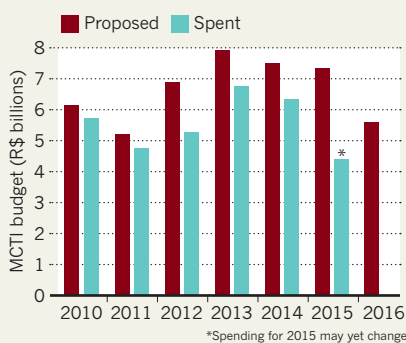
reagents and equipment. And promised government payments have been delayed. "Some institutions may not have received 50% of their funding yet," adds Barbosa.

Most foreign conference trips have been cancelled, says Vasco Azevedo, a geneticist at the Federal University of Minas Gerais in Belo Horizonte. And attendance at the Brazilian Society of Genetics annual meeting in Águas de Lindóia, São Paulo, this month — usually a lively gathering of 3,000 people — was down to less than half its usual level, he estimates. The funding situation also means that Brazil is unlikely to formally join the European Southern Observatory any time soon, Barbosa says, despite its Congress finally ratifying a 2010 agreement earlier this year.

State budgets — an increasingly important source of science funding in Brazil — are feeling the pinch as well. Most states' research funds come from a constitutionally mandated percentage of state revenues, which together amount to billions of reais each year. But many of the smaller, newer state agencies — which often rely on partnerships with embattled federal agencies — are reducing or postponing programmes, some by as much as 50%, says Sergio Gargioni, president of CONFAP, the umbrella organization for state funding

FEDERAL FUNDING WOES

The budget of the Brazilian Ministry of Science Technology and Innovation (MCTI) doubled from 2005 to 2010, but has fallen since the economy stalled in 2014.



UFES/LEI MARCELINO/REUTERS/CORBIS

SOURCE: MCTI

agencies, based in Brasilia.

Larger states are not immune. Rio de Janeiro's funding agency, which has seen its annual budget treble since 2006 to R\$450 million, has received less than 60% of its promised allocation so far this year, says Jerson Lima Silva, the agency's scientific director. And although Brazil's largest and most powerful state funding agency, the São Paulo Research Foundation (FAPESP), has received all of its mandated funds and has an endowment large enough to cover any shortfalls, it is still feeling the bite of shrinking state revenues and the poor exchange rate, says its scientific director, Carlos Henrique de Brito Cruz.

In an effort to revive science spending, science minister Aldo Rebelo is seeking a US\$2-billion loan from the Inter-American Development Bank, headquartered in Washington DC. Among other things, this will help to pay for planned National Science and Technology Institutes. Press reports suggest that the bank has agreed to this, but the MCTI told *Nature* that the deal requires approval from other government ministries and the Senate.

In a statement, a spokesman for the ministry said that Rebelo is also negotiating with the government to secure some of the 50% of oil-revenue funds for science that have yet to be allocated. "The MCTI has taken action in order to restore the budget and is seeking new sources of funding for the science and technology system," he said. For example, despite September cuts of R\$3.8 billion to a separate government fund called the Growth Acceleration Programme, Rebelo has secured funding for two cutting-edge science facilities that it was to have supported: a synchrotron light source set to come online in 2016 and a research nuclear-reactor facility due to open in 2018.

But with Brazil's economy showing no signs of recovery, many fear that the worst is yet to come. After further cuts were announced on 14 September, reports emerged that the government wants to merge the CNPq with CAPES, the Brazilian Federal Agency for the Support and Evaluation of Graduate Education, which is funded by the education ministry. The Brazilian Academy of Sciences and the Brazilian Society for the Advancement of Science, among others, came out strongly against such a move in a letter to the president, citing the agencies' "clear and complementary missions". And Gargioni thinks that the crisis could continue for years. "The situation is so black," says Azevedo. "We had very good graduate programmes and science, and now we will be back to 20 years ago. We need to do everything we can to save Brazilian science."

"We need to do everything we can to save Brazilian science."

MOLECULAR BIOLOGY

Bacteria yield new gene cutter

Smaller CRISPR enzyme should simplify genome editing.

BY HEIDI LEDFORD

The CRISPR/Cas9 gene-editing technique is revolutionizing genetic research: scientists have used it to engineer crops, livestock and even human embryos, and it may one day yield new ways to treat disease.

But one of the technique's pioneers thinks that he has found a way to make CRISPR even simpler and more precise. On 25 September, a team led by synthetic biologist Feng Zhang of the Broad Institute in Cambridge, Massachusetts, reported the discovery of a protein that may overcome one of CRISPR/Cas9's few limitations¹. Called Cpf1, the protein should make it easier to edit genes by replacing one DNA sequence with another, without compromising CRISPR's ability to disable genes.

The CRISPR/Cas9 system evolved as a way for bacteria and archaea to defend themselves against invading viruses. It is found in a wide range of these organisms, and uses an enzyme called Cas9 to cut DNA at a site specified by 'guide' strands of RNA. The cell's natural processes then repair the cuts. Researchers have now turned CRISPR/Cas9 into a molecular-biology powerhouse that can be used in other organisms.

CRISPR is much simpler than previous gene-editing methods, but Zhang thought that there was still room for improvement. So he and his colleagues searched the bacterial kingdom to find an alternative to the Cas9 enzyme commonly used in laboratories. In April, they reported² that they had discovered a smaller version of Cas9 in the bacterium *Staphylococcus aureus*. The small size makes the enzyme easier to shuttle into mature cells — a crucial destination for some potential therapies.

The team was also intrigued by Cpf1, a protein that looks very different from Cas9, but is present in some bacteria that use the CRISPR system. The scientists evaluated Cpf1 enzymes from 16 different bacteria, eventually finding two that could cut human DNA.

They also uncovered some curious variations in how Cpf1 and Cas9 work. Cas9 requires two RNA molecules to cut DNA; Cpf1 needs only one. The proteins also cut DNA at different places, offering researchers more options when selecting a site to edit, says epigeneticist Luca Magnani of Imperial College London.

And Cpf1 cuts DNA in a different way. Cas9

cuts both strands in a DNA molecule at the same position, leaving behind what molecular biologists call 'blunt' ends. But Cpf1 leaves one strand longer than the other, creating a 'sticky' end. Blunt ends are not as easy to work with because a DNA sequence could be inserted in either strand. But a sticky end will only pair with a complementary sticky end.

"The sticky ends carry information that can direct the insertion of the DNA," says Zhang, and that makes the process more controllable.

STICK WITH IT

Zhang's team is now working to use these sticky ends to improve the frequency with which researchers can replace a natural DNA sequence. Cuts left by Cas9 tend to be repaired by sticking the two ends back together, a process that can leave errors. Although it is possible that the cell will instead insert a designated new sequence at that site, that kind of repair occurs much less frequently. Zhang hopes that the unique properties of Cpf1 may be harnessed to make such insertions more frequent.

For Bing Yang, a plant biologist at the Iowa State University in Ames, this is the most exciting aspect of Cpf1. "Boosting the efficiency would be a big step for plant science," he says. "Right now, it is a major challenge."

Will the new enzyme surpass Cas9 in popularity? "It's too early to tell," says Zhang. "It certainly has some distinct advantages." The CRISPR/Cas9 system is so popular — and potentially lucrative — that it has sparked a fierce patent fight between the University of California, Berkeley, and the Broad Institute and its ally, the Massachusetts Institute of Technology in Cambridge. Zhang says that his lab will make the CRISPR/Cpf1 components available to academic researchers, as it has done with its CRISPR/Cas9 tools.

For now, the results stand as a testament that researchers still have more to learn from the genome-editing systems that bacteria have evolved. Microbiologist John van der Oost of Wageningen University in the Netherlands, who collaborated on the latest study with Zhang, plans to keep searching for new methods. "You never know whether one of these systems will be suitable for genome editing," he says. "There are still surprises ahead of us."

1. Zetsche, B. *et al. Cell* <http://dx.doi.org/10.1016/j.cell.2015.09.038> (2015).

2. Ran, F. A. *et al. Nature* **520**, 186–191 (2015).



BGI announced its plan to sell the micropigs as pets at a summit in Shenzhen, China.

GENE EDITING

Gene-edited pigs to be sold as pets

Chinese institute originally made the micropigs for research.

BY DAVID CYRANOSKI

Cutting-edge gene-editing techniques have produced an unexpected by-product — tiny pigs that a leading Chinese genomics institute will soon sell as pets.

BGI in Shenzhen, the centre that is famous for a series of high-profile breakthroughs in genomic sequencing, originally created the micropigs as models for human disease, by applying a gene-editing technique to a small breed of pig known as Bama. On 23 September, at the Shenzhen International Biotech Leaders Summit in China, BGI revealed that it would start selling the pigs as pets. The animals weigh about 15 kilograms when mature, or about the same as a medium-sized dog.

At the summit, the institute quoted a price tag of 10,000 yuan (US\$1,600) for the micropigs, but that was just to “help us better evaluate the market”, says Yong Li, technical director of BGI’s animal-science platform. In future, customers will be offered pigs with different coat colours and patterns, which BGI says it can set through gene editing.

With gene editing taking biology by storm, the field’s pioneers say that the application to pets was no big surprise. Some caution against

it: “It’s questionable whether we should impact the life, health and well-being of other animal species on this planet light-heartedly,” says geneticist Jens Boch at the Martin Luther University of Halle-Wittenberg in Germany. Boch helped to develop the technique used to create the pigs, which uses enzymes known as TALENs (transcription activator-like effector nucleases) to disable certain genes.

How to regulate the various applications of gene editing is an open question that scientists are already discussing with agencies across the world. BGI agrees on the need to regulate gene editing in pets as well as in the medical-research applications that make up the core of its micropig activities. Any profits from the sale of pets will be invested in this research. “We plan to take orders from customers now and see what the scale of the demand is,” says Li.

Compared to rats or mice, pigs are closer to humans physiologically and genetically, making them potentially more useful as a model organism for human disease. However, their larger size means that they cost more to keep and require bigger drug doses when they are used to test a pricey experimental medicine. Bama pigs, which weigh 35–50 kilograms (many farm pigs weigh more than 100 kilograms), have

previously been used in research.

To create the smaller, gene-edited micropigs, BGI made clones using cells taken from a Bama fetus. But before they started the cloning process, they used TALENs to disable one of two copies of the growth hormone receptor gene (*GHR*) in the fetal cells. Without the receptor, cells do not receive the ‘grow’ signal during development, resulting in stunted pigs.

SHOW STEALERS

BGI then created further micropigs by breeding stunted male clones with normal females. Only half of the resulting, naturally conceived offspring were micropigs, but the process is more efficient than repeating the full cloning procedure, and avoids potential health problems associated with cloning. Among the 20 second-generation gene-edited pigs, BGI has observed no adverse health effects, says Li.

He adds that the micropigs have already proved useful in studies of stem cells and of gut microbiota, because the animals’ smaller size makes it easier to replace the bacteria in their guts. They will also aid studies of Laron syndrome, a type of dwarfism caused by a mutation in the human *GHR* gene.

The decision to sell the pigs as pets surprised Lars Bolund, a medical geneticist at Aarhus University in Denmark who helped BGI to develop its pig gene-editing programme, but he admits that they stole the show at the Shenzhen summit. “We had a bigger crowd than anyone,” he says. “People were attached to them. Everyone wanted to hold them.”

In the United States, reports have surfaced of people who were disappointed when ‘teacup’ pigs weighing 5 kilograms grew into 50-kilogram animals. Gene-edited micropigs stay reliably small, the BGI team says. But gene editing will not solve other drawbacks of pet pigs, says Crystal Kim-Han, who runs a pig-rescue operation near Las Vegas, Nevada. If the animals are locked up in homes with no place to root or dig, they can become destructive. She also expects micropigs to have medical problems.

Some researchers think that dogs or cats will be next up for genetic manipulation. Scientists and ethicists agree that gene-edited pets are not that different from conventionally bred ones — the result is just achieved more efficiently. But that does not make the practice a good idea, says Jeantine Lunshof, a bioethicist at Harvard Medical School in Boston, Massachusetts, who describes both as “stretching physiological limits for the sole purpose of satisfying idiosyncratic aesthetic preferences of humans”.

Daniel Voytas, a geneticist at the University of Minnesota in Saint Paul, hopes that any buzz over gene-edited pets does not distract from or confuse efforts to use gene editing to alleviate human disease and create new crop varieties. “I just hope we establish a regulatory framework — guidelines for the safe and ethical use of this technology — that allows the potential to be realized,” he says. ■

ANCIENT TECHNOLOGY

Archimedes' fabled sphere brought to life

Curator recreates a 2,000-year-old model of the Universe.

BY JO MARCHANT

A mechanical model of the Universe attributed to the ancient Greek mathematician and polymath Archimedes has been reconstructed after more than two millennia. The metallic globe, which reproduces the motions of the Sun, Moon and planets across the night sky, is on display for the first time, at a museum in Basel, Switzerland.

The model, built by Michael Wright, a former curator at the Science Museum in London, is largely the product of erudite guesswork. But astrophysicist Mike Edmunds of Cardiff University, UK, says that it is a reminder that geared machines in antiquity were probably more complex than historians often assume.

Several ancient writers and poets describe mechanical models of the heavens¹, which they often attribute to Archimedes. The earliest and clearest of these appears in a dialogue² written by Roman author Marcus Tullius Cicero in the first century BC. One of Cicero's characters, Philus, describes how the Roman general Marcus Marcellus in 212 BC led an attack on Archimedes' home city of Syracuse (during which the mathematician was killed). As his troops ran the city, Marcellus took only one thing for himself: Archimedes' mechanical sphere.

When Philus later saw a demonstration of the device, he concluded that "the famous Sicilian had been endowed with greater genius than one

NATURE.COM
Nature Video: watch Wright's model of the sphere in action.
go.nature.com/wllt41



Michael Wright's machine models the heavens.

would imagine it possible for a human being to possess". Solid globes marked with star constellations were common at the time. But Archimedes' invention, Philus notes, also included the Sun, Moon and the five known planets, displaying as it turned "those various and divergent movements with their different rates of speed".

Historians once thought that Cicero's description was fabricated or exaggerated. But studies of a relic known as the Antikythera mechanism, found on a shipwreck from the first century BC, have changed that view. The device turned out to be a clockwork calendar that could model the movements of celestial bodies and predict solar and lunar eclipses — and thus proved that

complex geared astronomical devices did exist in antiquity; it consisted of more than 30 bronze gearwheels inside a wooden box the size of a phone book (see *Nature* **444**, 534–538; 2006).

Most specialists have concluded that Cicero was describing a similar machine. But Wright — who has previously built two working models of the Antikythera mechanism — points out that descriptions of Archimedes' device use the Latin word *sphaera* (*sphaira* in Greek). "The Antikythera mechanism is not a sphere; it's a shoebox," he says.

Other scholars counter that 'sphere' could have been a generic term for astronomical models, regardless of their shape. But Wright retorts that in Cicero's description, when the globe turned, "the Moon was always as many revolutions behind the Sun on the bronze contrivance as would agree with the number of days it was behind it in the sky". This implies that the device turned once each day, he says, which makes no sense for a flat dial.

Wright built his machine with similar techniques to those that Archimedes might have used. He engraved pictures of the Greek constellations on the surface of the 20-centimetre-wide globe and mounted it in a wooden box, which hides the portion of the globe below the horizon at any given time.

As the globe is turned by hand, 24 gearwheels hidden inside drive curved pointers on the outside. Those marking the Sun and the Moon move at a constant speed, whereas the planets meander, moving back and forth with respect to the fixed stars, just as in the real sky.

No one knows whether Archimedes truly came up with such a device, but Wright argues that he was perfectly positioned to do so. The ancient scholar was a brilliant mathematician and famous for building ingenious machines.

The model is at the Basel Museum of Ancient Art and Ludwig Collection, as part of an exhibition of artefacts from the Antikythera wreck. ■

1. Edmunds, M. G. *Contemp. Phys.* **55**, 263–285 (2014).
2. Cicero, M. T. *De Re Publica* Vol. 213 (transl. Keyes, C. W.) 40–44 (Loeb, 1928).

ADAM LEVY/NATURE

NUMBER THEORY

Maths whizz solves a master's riddle

Terence Tao builds on an online collaboration to attack the Erdős discrepancy problem.

BY CHRIS CESARE

A maths puzzle that resisted solution for more than 80 years — including computerized attempts to crack it — seems to have yielded to a single mathematician.

On 17 September, Terence Tao, a mathematician at the University of California, Los

Angeles, whose body of work earned him the prestigious Fields Medal in 2006, submitted a paper to the arXiv preprint server claiming to prove a number-theory conjecture posed by mathematician Paul Erdős in the 1930s (T. Tao. Preprint available at <http://arxiv.org/abs/1509.05363>; 2015).

"Terry Tao just dropped a bomb," tweeted

Derrick Stolee, a mathematician at Iowa State University in Ames.

Like many puzzles in number theory, the Erdős discrepancy problem is simple to state but devilishly difficult to prove. Erdős, who died in 1996, speculated that any infinite string made up of the numbers 1 and –1 could add up to an arbitrarily large (positive or negative) ▶

► value by counting only the numbers at a fixed interval for a finite number of steps.

Tao's proof shows that these sums can, in fact, grow infinitely large for any arbitrary sequence, although it does not provide a way to calculate their value for a given instance.

The proof has not yet undergone rigorous peer review, but experts have expressed no concern over whether it will survive a critical look. "I'm completely confident," says Gil Kalai, a mathematician at the Hebrew University of Jerusalem.

Tao's proof comes after years of attempts to solve the problem by hand and computer. The most recent campaign began in December 2009 and gathered steam in 2010. Mathematician Tim Gowers at the University of Cambridge, UK, suggested focusing on Erdős's problem for the fifth PolyMath Project, an online collaboration in which mathematicians work together on a single mathematical puzzle. Tao was one of several dozen participants.

The effort fizzled out in 2012, but participants did manage to show that proving the conjecture for a certain family of sequences was good enough to prove it in general. That family has arbitrary 1s and -1s only in the spots indexed by prime numbers.

In February 2014, researchers presented a computer proof for a special case of the conjecture: they showed that it is always possible



Terence Tao solved a major number-theory puzzle after being inspired by a comment on his blog.

to find a sum that is bigger than 2 (B. Konev and A. Lisitsa Preprint available at <http://arxiv.org/abs/1402.2184>; 2014). However, they failed to prove that there is always a sum bigger than 3. Tao's proof demonstrates that there is always a sum bigger than any finite number.

No one else managed to make major progress after the computational attempt. Tao had been working on a different problem early in September, when a timely comment on his blog suggested that the problem might be related to the Erdős conjecture. "At first, I

thought the connection was only superficial," says Tao. But he soon realized that combining the commenter's fresh insight with previous results could lead to a solution. He submitted his paper less than two weeks later and included an acknowledgement thanking the commenter, Uwe Stroinski, a maths instructor in Reutlingen, Germany, who holds a PhD in mathematics from the University of Tübingen.

Tao has submitted his proof to the open-access journal *Discrete Analysis*, run by Gowers. The journal, which was founded in early September, provides conventional peer review but accepts only papers that have already been posted on arXiv, thereby avoiding major publishing costs. "Tim's journal is a promising experiment in completely open-access publishing," says Tao.

Erdős, who wrote a letter recommending Tao for admission to Princeton University in New Jersey, often offered cash prizes for solving the problems he posed. He set the prize for the discrepancy problem at US\$500. Since his death, others have taken it upon themselves to award those prizes on his behalf.

When asked whether he would accept the prize were someone to offer it, Tao demurred. "It was traditional to not actually cash the prizes that Erdős did award while he was alive," he says. "People usually framed the cheque instead." ■

RICHARD HARTOG/LOS ANGELES TIMES/GETTY



BITCOIN AND BEYOND

The digital currency has caused any number of headaches for law enforcement. Now entrepreneurs and academics are scrambling to build a better version.

BY ANDY EXTANCE

When the digital currency Bitcoin came to life in January 2009, it was noticed by almost no one apart from the handful of programmers who followed cryptography discussion groups. Its origins were shadowy: it had been conceived the previous year by a still-mysterious person or group known only by the alias Satoshi Nakamoto¹. And its purpose seemed quixotic: Bitcoin was to be a 'cryptocurrency', in which strong encryption algorithms were exploited in a new way to secure transactions. Users' identities would be shielded by pseudonyms. Records would be completely decentralized. And no one would be in charge — not governments, not banks, not even Nakamoto.

Yet the idea caught on. Today, there are some 14.6 million Bitcoin units in circulation. Called bitcoins with a lowercase 'b', they have a collective market value of around US\$3.4 billion. Some of this growth is attributable to criminals taking advantage of the anonymity for drug trafficking and worse. But the system is also drawing interest from financial institutions such as JP Morgan Chase, which think it could streamline their internal payment processing

and cut international transaction costs. It has inspired the creation of some 700 other cryptocurrencies. And on 15 September, Bitcoin officially came of age in academia with the launch of *Ledger*, the first journal dedicated to cryptocurrency research.

What fascinates academics and entrepreneurs alike is the innovation at Bitcoin's core. Known as the block chain, it serves as the official online ledger of every Bitcoin transaction, dating back to the beginning. It is also the data structure that allows those records to be updated with minimal risk of hacking or tampering — even though the block chain is copied across the entire network of computers running Bitcoin software, and the owners of those computers do not necessarily know or trust one another.

Many people see this block-chain architecture as the template for a host of other applications, including self-enforcing contracts and secure systems for online voting and crowdfunding. This is the goal of Ethereum, a block-chain-based system launched in July by the non-profit Ethereum Foundation, based in

Baar, Switzerland. And it is the research agenda of the Initiative for CryptoCurrencies and Contracts (IC3), an academic consortium also launched in July, and led by Cornell University in Ithaca, New York.

Nicolas Courtois, a cryptographer at University College London, says that the Bitcoin block chain could be "the most important invention of the twenty-first century" — if only Bitcoin were not constantly shooting itself in the foot.

Several shortcomings have become apparent in Bitcoin's implementation of the block-chain idea. Security, for example, is far from perfect: there have been more than 40 known thefts and seizures of bitcoins, several incurring losses of more than \$1 million apiece.

Cryptocurrency firms and researchers are attacking the problem with tools such as game theory and advanced cryptographic methods. "Cryptocurrencies are unlike many other systems, in that extremely subtle mathematical bugs can have catastrophic consequences," says Ari Juels, co-director of IC3. "And I think when weaknesses surface there will be a need to appeal to the academic community where the relevant expertise resides."

THE BITCOIN GAME

The cryptocurrency's **mining** process is designed to produce a secure online ledger of every Bitcoin transaction — even though no one is in charge.

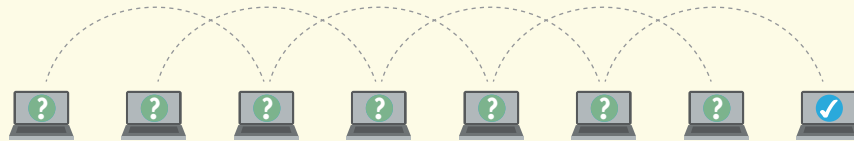
THE TRANSACTION

Bob sends some bitcoins to Alice, both use pseudonyms to keep their identities secret.



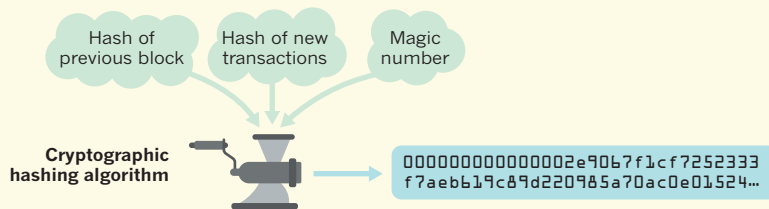
THE MINERS

Digital copies of the transaction are passed to **miners** for verification. The miners are individuals or groups running the **Bitcoin software** in a worldwide network of independent computers. They compete to turn the latest transactions into a **block**. Roughly every ten minutes, one of them succeeds.



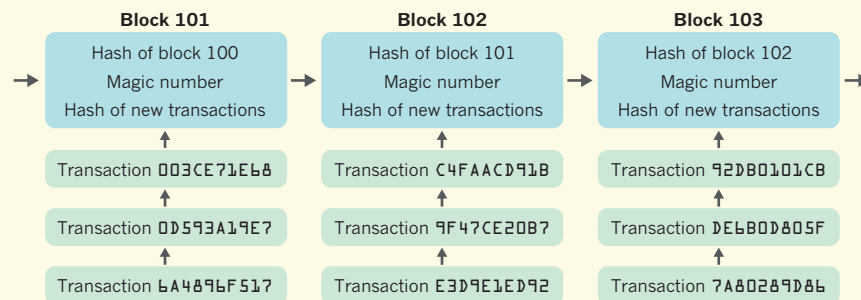
THE WINNING BLOCK

Encrypting the transactions creates a **hash** — a seemingly random sequence of numbers and letters. The miners try to find a **magic number** that when encrypted alongside the transactions and the most recent block in the chain creates a hash that starts with a particular number of zeros. Although this number is very hard to find, once a solution has been found it can be verified easily by the other miners. The first miner to solve the problem is rewarded with bitcoins, and the block is added to the **block chain**.



THE BLOCK CHAIN

The block chain is an online ledger that records every Bitcoin transaction ever made. A copy of the block chain is held by each miner and it is used as proof of ownership for all bitcoins. Chronological order is very important in the chain. If Bob has already spent his bitcoins elsewhere it will be recorded in the block chain and his transfer to Alice will be rejected.



Academic interest in cryptocurrencies and their predecessors goes back at least two decades, with much of the early work spearheaded by cryptographer David Chaum. While working at the National Research Institute for Mathematics and Computer Science in Amsterdam, the Netherlands, Chaum wanted to give buyers privacy and safety. So in 1990 he founded one of the earliest digital currencies, DigiCash, which offered users anonymity through cryptographic protocols of his own devising.

DigiCash went bankrupt in 1998 — partly

because it had a centralized organization akin to a traditional bank, yet never managed to fit in with the financial industry and its regulations. But aspects of its philosophy re-emerged ten years later in Nakamoto's design for Bitcoin. That design also incorporated crowdsourcing and peer-to-peer networking — both of which help to avoid centralized control. Anyone is welcome to participate: it is just a matter of going online and running the open-source Bitcoin software. Users' computers form a network in which each machine is home to one

constantly updated copy of the block chain.

Nakamoto's central challenge with this wide-open system was the need to make sure that no one could find a way to rewrite the ledger and spend the same bitcoins twice — in effect, stealing bitcoins. His solution was to turn the addition of new transactions to the ledger into a competition: an activity that has come to be known as mining (see 'The Bitcoin game').

Mining starts with incoming Bitcoin transactions, which are continuously broadcast to every computer on the network. These are collected by 'miners' — the groups or individuals who choose to participate — who start competing for the right to bundle transactions into a new block. The winner is the first to broadcast a 'proof of work' — a solution showing that he or she has solved an otherwise meaningless mathematical puzzle that involves encrypted data from the previous block, and lots of computerized trial and error. The winning block is broadcast through the Bitcoin network and added to the block chain, with the proof of work providing an all but unbreakable link. The block chain is currently almost 400,000 blocks long.

In principle, this competition keeps the block chain secure because the puzzle is too hard for any one miner to solve every time. This means that no one will ever gain access to the encrypted links in the block chain and the ability to rewrite the ledger.

Mining is also a way to steadily increase the bitcoin supply: the miner who wins each block gets a reward, currently 25 new bitcoins. That is worth almost \$6,000 at today's prices. Nakamoto's design controls the supply increase by automatically adjusting the difficulty of the puzzle so that a new block is added roughly every ten minutes. In addition, the reward for creating a block decreases by half roughly every four years. The goal is to limit the supply to a maximum of 21 million bitcoins.

The network cannot determine the value of bitcoins relative to standard currencies, or real-world goods and services. That has been left to market forces, with people trading bitcoins on online exchanges. One result is that the market price has gyrated spectacularly — especially in 2013, when the asking price soared from \$13 per bitcoin in January to around \$1,200 in December. That would have made the first real-world products ever paid for with the cryptocurrency — a pair of Papa John's pizzas, purchased for 10,000 bitcoins on 22 May 2010 — worth almost \$12 million.

PUZZLE SOLUTIONS

It did not take long for the problems with Bitcoin to become apparent. For example, because users are allowed to mask their identity with pseudonyms, the currency is perfect for screening criminal activity. That was behind the success of the online black market Silk Road, which the FBI shut down in 2013; its founder was sentenced to life in prison in May this year. But Bitcoin also had a key role

in funding the whistle-blowing website WikiLeaks — an outcome that some would call beneficial. It is difficult for society to work out a legal framework to differentiate between good and bad uses of this technology, says Arvind Narayanan, a computer scientist at Princeton University in New Jersey. “How do you regulate around Bitcoin without banning the technology itself?” he asks.

Other issues surfaced with Bitcoin’s mining procedure. As the currency has gained value, for example, mining competition has become fiercer, with increasingly specialized computers solving the puzzles ever faster. Courtois, who has found ways to streamline the puzzle-solving process², says that at one point he was successfully earning \$200 a day through mining. The rivalry has driven the establishment of large Bitcoin-mining centres in Iceland, where cooling for the computers is cheap. According to one estimate from 2014, Bitcoin miners collectively consumed as much power as the whole of Ireland³.

WORKING TOGETHER

Intensified Bitcoin mining has also led individual miners to pool their computational resources. Last year, the largest mining pool, GHash.IO, briefly exceeded 50% of total Bitcoin mining power — which is problematic because anyone who controls more than half of the mining power could start beating everyone else in the race to add blocks. This would effectively give them control of the transaction ledger and allow them to spend the same bitcoins over and over again. This is not just a theoretical possibility. Successful ‘51% attacks’ — efforts to dominate mining power — have already been mounted against smaller cryptocurrencies such as Terracoin and Coinedcoin; the latter was so badly damaged that it ceased operation.

To reduce the threat from mining pools, some existing cryptocurrencies, such as Litecoin, use puzzles that call more on computer memory than on processing power — a shift that tends to make it more costly to build the kind of specialized computers that the pools favour. Another approach, developed by IC3 co-director Elaine Shi and her collaborators⁴, enlists a helpful kind of theft. “We are cryptographically ensuring that pool members can always steal the reward for themselves without being detected,” explains Shi. Their supposition is that miners would not trust each other enough to form into pools if their fellow pool members could easily waltz off with the rewards without sharing. They have built a prototype of the algorithm, and are hoping to see it tested in Bitcoin and other cryptocurrencies.

Another problem is the profligate amount of electricity used in Bitcoin mining. To reduce wastage, researchers including Shi and Juels have proposed a currency called Permacoin⁵. Its proof of work would

require miners to create a distributed archive for valuable data such as medical records, or the output of a gene-sequencing centre. This would not save energy, but would at least put it to better use.

The security of cryptocurrencies is another huge concern. The many thefts of bitcoins do not result from the block-chain structure, says Narayanan, but from Bitcoin’s use of standard digital-signature technology. In digital signatures, he explains, people have two numeric keys: a public one that they give to others as an address to send money to, and a private one that they use to approve transactions. But the security of that private key is only as good as the security of the machine that stores it, he says. “If somebody hacks your computer, for example, and steals your private keys, then essentially all of your bitcoins are lost.”

Security is such a concern for consumers that Narayanan thinks Bitcoin is unlikely to find widespread use. So his team is working on a better security scheme that splits private keys across several different devices, such as an individual’s desktop computer and smartphone, and requires a certain proportion of the fragments to approve a payment⁶. “Neither reveals their share of the key to each other,” says Narayanan. “If one machine gets hacked, you’re still OK because the hacker would need to hack the others to steal your private key. You’ll hopefully notice the hack happened before they have the chance.”

Other thefts have occurred because the private key needs to be combined with a random number to create a transaction signature. Some software — such as Bitcoin apps developed for Android smartphones — has generated random numbers improperly, making them easier to guess. This has allowed hackers to steal somewhere between several thousand and several million dollars’ worth of bitcoins, says Courtois, who has been investigating such vulnerabilities⁷. “It’s embarrassing,” admits David Schwartz, chief cryptographer at cryptocurrency developer Ripple Labs in San Francisco, California. “We as an industry just seem to keep screwing up.”

INTO THE ETHER

The block chain is a remarkably powerful idea that could be applied to much more than just transaction records, says Gavin Wood, co-founder of Ethereum and chief technology officer of its foundation. One use might be to develop computerized, self-enforcing contracts that make a payment automatically when a task is complete. Others might include voting systems, crowdfunding platforms, and even other cryptocurrencies. Wood says that Ethereum is best used in situations for which central control is a weakness — for example, when users do not necessarily trust one another. In 2014, to make it easier to develop such applications, Wood and fellow programmer Vitalik Buterin devised a way to combine

the block chain with a programming language. Ethereum raised 30,000 bitcoins through crowdfunding to commercialize this system.

To prevent the basic cryptography-related mistakes that have plagued Bitcoin, Ethereum has recruited academic experts to audit its protocol. Shi and Juels are looking for ways that Ethereum could be abused by criminals⁸. “The technology itself is morally neutral, but we should figure out how to shape it so that it can support policies designed to limit the amount of harm it can do,” says Juels.

Like Bitcoin, Ethereum is not under anyone’s direct control, so it operates outside national laws, says Wood. However, he adds that technologies such as music taping and the Internet were also considered extralegal at first, and seemed threatening to the status quo. How Bitcoin, Ethereum and their successors sit legally is therefore “something that, as a culture and society, we’re going to have to come together to deal with”, he says.

Juels suspects that Bitcoin, at least, will not last as an independent, decentralized entity. He points out how music streaming has moved from the decentralized model of peer-to-peer file-sharing service Napster to commercial operations such as Spotify and Apple Music. “One could imagine a similar trajectory for cryptocurrencies: when banks see they’re successful, they’ll want to create their own,” he says.

Courtois disagrees. He calls Bitcoin “the Microsoft of cryptocurrency”, and maintains that its size and dominance mean that it is here to stay. As soon as any new innovations come along, he suggests, Bitcoin can adopt them and retain its leading position.

Whatever the future holds for Bitcoin, Narayanan emphasizes that the community of developers and academics behind it is unique. “It’s a remarkable body of knowledge, and we’re going to be teaching this in computer science classes in 20 years, I’m certain of that.” ■

Andy Exantse is a freelance writer in Exeter, UK.

1. Nakamoto, S. *Bitcoin: A Peer-to-Peer Electronic Cash System* (2008); available at <https://bitcoin.org/bitcoin.pdf>
2. Courtois, N. T., Grajek, M. & Naik, R. Preprint available at <http://arxiv.org/abs/1310.7935> (2013).
3. O’Dwyer, K. J. & Malone, D. *25th IET Irish Signals & Systems Conf. 2014 and 2014 China-Ireland Int. Conf. on Information and Communities Technologies* 280–285 (2014).
4. Miller, A., Shi, E., Kosba, A. & Katz, J. *ACM Conf. Computer and Communications Security* (2015); preprint available at <http://go.nature.com/2i2sfe>
5. Miller, A., Juels, A., Shi, E., Parno, B. & Katz, J. *IEEE Symp. Security and Privacy* 475–490 (2014).
6. Goldfeder, S. *et al. Securing Bitcoin Wallets via a New DSA/ECDSA Threshold Signature Scheme* (2015); available at <http://go.nature.com/rnqp4q>
7. Courtois, N. T., Emirdag, P. & Valsorda, F. *Cryptology ePrint Archive Report 2014/088* (2014).
8. Juels, A., Kosba, A. & Shi, E. *The Ring of Gyges: Using Smart Contracts for Crime* (2015); Preprint available at <http://go.nature.com/sbsdqk>

NATURE.COM
To hear more about
cryptocurrencies:
go.nature.com/icc8kv

MOUNTAIN BATTLE

Plans to build one of the world's biggest telescopes on Mauna Kea in Hawaii are mired in conflict. Four people involved in the fight explain their diverse views.

BY ALEXANDRA WITZE

A low, mournful note rings across the broad summit of Mauna Kea, the highest peak in Hawaii. Joshua Lanakila Mangauil lowers his conch-shell horn and begins to walk over the craggy volcanic rock. Behind him come a dozen other Native Hawaiian men and women, many carrying the red, white and blue state flag. Together, they sing traditional chants as they hike up a volcanic ridge, headed for the top of the mountain.

On this July day, Mangauil is leading a reverent but short visit to the 4,200-metre-high summit. Native Hawaiian tradition calls for visitors to pay their respects to the sacred *mauna*, or mountain, and then leave. The group planned to stay a few hours, then get back into its pickup trucks and drive to a camp farther down. There they would resume the task that has consumed Mangauil and others for the greater part of a year: protecting the *mauna* from an effort to build a massive telescope.

An international consortium plans to construct the Thirty Meter Telescope (TMT) on top of Mauna Kea on Hawaii's Big Island. A cutting-edge astronomical facility, it would have a light-gathering mirror 3 times bigger than any of the 13 other telescopes already on the mountain, which include some of the largest and most scientifically productive observatories in the world.

TMT construction began in April and stopped almost immediately when demonstrators led by Mangauil blocked the vehicles from reaching the summit. They say that the TMT would violate both a fragile ecosystem and indigenous rights that have not been properly valued by astronomers. "Before you look into space, you need to respect this place," Mangauil says.

The battle that erupted this year echoes previous clashes. Native Hawaiians, environmentalists, scientists and other interest groups have wrangled for decades over the environmental and cultural impacts of the summit telescopes (see 'Peak passions'). But the fight over the TMT has grown larger and more divisive than past debates. Thanks to a renaissance of Native Hawaiian pride and the flashpoint potential of social media, the protests have tapped into broader anger against the US government and its past behaviour towards the islands and their native peoples.

Mangauil is adamant that the telescope should not be built — but others have equally passionate, divergent perspectives. Here, *Nature* talks to four people living in Hawaii and involved in the controversy, whose differing viewpoints reveal the complexity of finding common ground and securing the future of astronomy there.

KENT NISHIMURA

Joshua Lanakila Mangauil

THE OPPONENT

When Mangauil is not paying tribute on Mauna Kea's summit, he can often be found about 1,400 metres below, at a stopping point that includes a visitor's centre and a makeshift camp set up by demonstrators in March. As one of the leaders of the protest group there, Mangauil frequently talks to supporters, tourists, journalists — anyone who stops by.

In late July, the camp consisted of a small thatched hut, of traditional design, in which a Native Hawaiian man wearing a loincloth and shoulder cape was waiting to describe cultural customs to any visitors. Half a dozen other protesters stood around talking to each other quietly. Next to the hut, in a large tarpaulin enclosure amid camping beds and food supplies, Mangauil was working on his mobile phone, wearing a T-shirt and hoodie. He broke away to talk to a group of children who had travelled from the neighbouring island of Maui to see him. Many wore the bright-red *kū kia'i mauna*, or 'guardians of the mountain', T-shirts that have become popular among TMT protestors.

Mangauil bent over to greet several of the students, touching his nose to theirs in the traditional manner that symbolizes an exchange of breath. "Our creation is connected to this mountain," he explained, as some of the kids recorded him on their smartphones. In Hawaiian history, the sky father and the earth mother came together to create the Big Island, and

Mauna Kea is the centre. "This mountain is the oldest sibling that watches over all of us," Mangauil said. "It collects the clouds, it channels the water, it gives us life."

Mangauil has felt this connection since his childhood in a rainy town beneath Mauna Kea's northern slopes. "The river behind my house comes from this place," he says. "This has always been my mountain."

Like many in the younger generation of Native Hawaiians — Mangauil is 28 — he attended a Hawaiian-language immersion school. Such institutions have helped to reinvigorate Hawaiian culture after a long period during which it was suppressed. In Mangauil's grandparents' generation, children at school were beaten if they spoke their native language, and heard more about George Washington than about their own Kamehameha the Great, who united the Hawaiian islands two centuries ago. But beginning in the 1970s, activists began to push back. Now, Hawaiian students can choose from a variety of immersion programmes in public and charter schools across the islands.

Mangauil says that he was too young to be involved when the TMT began the seven-year process of obtaining state permits to build on Mauna Kea, but one of his teachers was a long-time activist fighting development on the mountain. Through her, he grew familiar with the issues surrounding Mauna Kea. After graduating, he returned to his school to work as a teacher.

As his interest in cultural matters grew, Mangauil set up a business to consult on Native

Hawaiian issues and began spending more time on Mauna Kea. In October 2014, when TMT officials organized a groundbreaking ceremony with visiting dignitaries, Mangauil surprised them — and himself — by jumping in front of the cameras and denouncing the project. "That was not planned," he says. "I was upset."

Six months later, Mangauil was again in the front of a protest group blocking the path of TMT construction trucks. He and 30 other demonstrators were arrested, booked and released. (A wealthy descendant of Hawaii's monarchy has put up much of the bail money.)

Now Mangauil spends most of his time in the role of mountain guardian: leading protests, testifying at hearings and travelling to other islands to meet with activists. He helps to ignite demonstrations through his popular Facebook pages. He is also dipping his toe into politics, by running for a seat in a newly formed group that aims to build a governing base for a future Hawaiian nation. (Hawaii's monarchy was overthrown in 1893 by pro-US interests; it became a US state in 1959.)

Mangauil's more-immediate goal is simple: to stop the TMT from being built on Mauna Kea. "We are fighting for the rights of the mountain," he says. "I have nothing against astronomy — just don't put it up there."

Looking ahead, Mangauil sees a day when astronomers will leave Mauna Kea. The existing telescopes are legally allowed to operate there until their lease to the site expires in 2033. At that point, Mangauil argues, they should all be dismantled. "Then the mountain can rest."

PEAK PASSIONS

With some of the best astronomical viewing conditions in the world, Mauna Kea in Hawaii has long played host to top telescopes. But scientists have frequently clashed with Native Hawaiians and environmentalists over facilities on the mountain.

1960

A tsunami devastates the city of Hilo. Looking for ways to rebuild the economy, local businessmen begin recruiting astronomers to develop observatories there.

1968

The state of Hawaii gives the University of Hawaii a 65-year lease to operate the summit area of Mauna Kea as a science reserve.

1968

The US Air Force builds the first research telescope on Mauna Kea, a 0.6-metre facility.



Alexis Acohido

THE SUPPORTER

As Mangauil sounds his conch on Mauna Kea's summit, Alexis Acohido stands nearby trying to explain her feelings about the protest movement. "I'm conflicted," she says finally. "I'm really conflicted."

Acohido, aged 22 and part Native Hawaiian, is of Mangauil's generation but not his mindset. Growing up on the island of Oahu near the heart of Honolulu, Acohido was always drawn towards science. "In high school, I wanted to

be a biologist, but when I got to college I had all these math credits and thought, why don't I get my degree in math?" she says.

Acohido first heard of the TMT several years ago, while sitting in an orientation lecture for a summer astronomy internship. A project scientist spoke about how the TMT would see stars and galaxies with unprecedented clarity, better even than today's views from the Hubble Space Telescope. "What they had planned sounded

really awesome," she says. "I thought it would be cool to have for Hawaii."

So when TMT protests began to spread this spring, she decided to become more active. In a debate during a philosophy class, she spoke out in favour of the TMT, and a fellow student asked her to write an opinion piece for the university's newspaper. After that appeared, the public-relations firm that has been handling TMT affairs asked her to expand on her ideas for Honolulu's major newspaper, and so she published a commentary there in April.

Acohido argued that the TMT should be built and that it would bring opportunities to students in Hawaii, a state that has typically ranked below US averages in school performance. (Among other contributions, the TMT has set up an educational fund that awards US\$1 million annually to Big Island students in technical fields.) Her article did not go down well. Many people told her "How could you? You need to get back to your roots", she says. Some people who commented online claimed she had no cultural authority to speak about Native Hawaiian issues and had been brainwashed by TMT leaders. "I've been called a bad Hawaiian so much it's not funny," she says.

Acohido says that she even faced opposition from some members of her family. But others have offered encouragement, including her Native-Hawaiian grandmother. "She's been super-supportive," Acohido says. "She'll always save any news piece that comes out, for me to read when I get home."

In many ways, Acohido represents the next generation of Hawaiian scientists that the TMT and other observatories hope to foster. She graduated with her mathematics degree earlier this year and now works as a communications intern in the Hilo offices of Gemini Observatory, which runs an 8-metre telescope on Mauna Kea. The day she saw Mangauil was her first time at the summit. She handed over her smartphone so she could get a photo of herself with the gleaming Gemini dome behind her.

Watching the demonstrators, Acohido talked about the conflict she feels. "It's important to uphold your cultures and traditions, but I also think it's important to pick your battles," she says. "A lot of their anger is misplaced." In her view, they should focus their wrath on the University of Hawaii (UH), which has managed the mountain observatories since the 1960s. "If they are going to be mad at anyone," she says, "they should be mad at UH."

KENT NISHIMURA



1990s

The biggest telescopes on Mauna Kea start operations: the 8-metre Gemini North and Subaru, and the twin 10-metre Kecks. Environmentalists, Native Hawaiians and others argue that the mountain has been desecrated.

2006

Following strong opposition and legal setbacks, NASA decides not to build supplementary 'out-rigger' telescopes at Keck.

2009

Officials for the Thirty Meter Telescope (TMT) announce their plan to build a next-generation observatory on Mauna Kea.

2015

Native Hawaiian protestors halt TMT construction. The state's governor announces that the university must decommission as many telescopes as possible.

THE BATTLE-SCARRED ASTRONOMER

The UH's Institute for Astronomy sits above the urban bustle of Honolulu in a neighbourhood of quiet winding streets. Inside, on a muggy August morning, astronomer Bob McLaren sighs at the current chaos engulfing Hawaiian astronomy. "The core disagreement is simple, but there's no easy solution," he says.

McLaren knows these battles all too well. As the university representative tasked with developing astronomy facilities, he has taken part in some of the most contentious fights over the future of Mauna Kea.

He came to Hawaii from Canada's University of Toronto in 1982, drawn to Mauna Kea as the best site to do infrared astronomy. Observing variable stars with the 4-metre Canada-France-Hawaii Telescope, he helped to recalibrate the distances to many nearby galaxies. But as larger telescopes started to come online, McLaren saw a role in helping to manage how astronomy was conducted on the mountain. He moved into his UH administration job and onto the front lines of public battles.

Some of his most painful professional memories are from the 1990s, when university and state officials were working to adopt a master plan for the mountain's future. In a blistering 1998 report, auditors slammed the university for failing to balance telescope development with Mauna Kea's archaeological, cultural and environmental resources. Among its criticisms, the assessment noted that Native Hawaiian cultural practitioners — generally older men and women who travel to the mountain's summit to carry out personal devotions — charged that rubbish and development had desecrated it.

Relations between astronomers and Native Hawaiians deteriorated so much that the late Senator Daniel Inouye, a legendary figure in Hawaiian politics, had to intervene. He forced nine people from each side, including McLaren as the institute's associate director, to sit down and talk out their differences. "It was kind of awkward at first, but it actually worked," McLaren says. "We could discuss why we thought certain things and why they thought certain things, and what we were going to do about it. But it took time." Those 1999 talks helped to get a comprehensive plan for the mountain approved. And some of the Native Hawaiians on the panel formed a cultural advisory council that provides input into the management of Mauna Kea. It was a rare example of people with different interests managing to have a productive dialogue about the mountain's future, McLaren says.



Bob McLaren

Despite all the conversations, however, people were uncomfortable with the concept of future development, so the final master plan failed to lay out a clear sense of whether and how big telescopes could be built on the mountain. "We weren't able to achieve that," says McLaren. "That was a bit disappointing." In 2006, in the wake of the master-plan controversy, associated lawsuits and strong opposition, NASA pulled funding for a project that would have added up to six 'outrigger' telescopes to the twin 10-metre Keck telescopes, currently the largest atop Mauna Kea.

Even though the master plan did not go as far as McLaren had wanted, he says that it has helped to shape the current project. When the TMT team decided in 2009 to build on Mauna Kea, it worked within the guidelines of the plan to minimize the telescope's impact on the mountain. Physically, the dome is slated to sit about 150 metres below the summit ridge, making it less prominent. Project officials consulted with a number of Native Hawaiian groups, including the Mauna Kea cultural advisory council, and made plans to limit traffic to the summit and have local voices deeply involved in all stages of construction. The TMT is also the first Mauna Kea telescope with more than a token rent; it will pay \$1 million

a year for its space, with most of that flowing directly to mountain stewardship.

That made it all the more surprising — at least to astronomers — when Mangauil and others jumped in front of the cameras at the groundbreaking ceremony last October. "I can't explain what suddenly got all of these new people involved," says McLaren. In the past, "there were elements of this sovereignty movement and disenfranchisement in there, but they were secondary", he says. "Now it's a lot more complicated. Many of the people we're hearing speak would like to have a society that feels more Hawaiian to them, in the spirit of things they've been taught in school."

McLaren is frustrated by what he sees as changing cultural values among all the groups with a stake in the mountain. "I'd like people to tell me, if they got in a time machine and went back to 1964, what they would have done differently," he says. "Don't just criticize what's on the mountain now — tell us what we did wrong, in the context of what actually happened."

Most days, McLaren is optimistic that the TMT will be built in Hawaii. Other days, he cannot quite see a way through the conflict. "People like me get a little cynical," he says. "We've seen this movie before."



KENT NISHIMURA

THE BRIDGE BUILDER

Doug Simons

The discord might be familiar, but Doug Simons refuses to give up hope.

Simons has worked in Hawaii for three decades, including stints directing Gemini and now the Canada-France-Hawaii Telescope. He specialized in developing instruments to study the Universe in infrared wavelengths, and is now thinking of ways in which the diverse observatories on Mauna Kea can work together more closely.

Simons's ties to the mountain go far beyond his work. As someone who hunts birds on Mauna Kea, he wants to see its environment preserved and he has even made plans to have his ashes scattered nearby. "It tears me up to see my community being torn apart," he says. "At the end of the day, it's not me the astronomer, it's me the Big Island resident that has made me commit to finding a solution for my neighbours."

So Simons has been meeting with anyone who wants to talk to him: Native Hawaiians whose families are divided, businessmen who wonder whether Hawaii will invest in high-tech industries, secondary-school students who desire local well-paying technical jobs and those who want the telescopes removed. He does it the old-fashioned way, one face-to-face sit-down at a time. "I'm kind of old-timey," he says. "I'm not a Facebook guy."

Sometimes these conversations happen by chance: Simons recently ran into Mangauil at an airport and the two shared a beer. The discussions are always deeply personal. "You

have to come out of your comfort zone as a scientist and get into the emotional arena to make the connection," he says. When speaking with Native Hawaiians about Mauna Kea, he opens up about his quiet Catholic faith, his daily prayers. "It's my spiritual component trying to map out to theirs," he says.

He hopes and believes that these conversations will make a difference. "There is no way to make everybody happy on the mountain," he says. "Historically, these things have been worked through some sort of give-and-take process — I don't see why that can't happen here."

Some compromises have already been found. The UH is working through long-held plans to decommission some of the older telescopes on Mauna Kea and return the land to its natural state. Two are slated to be removed soon, and a third demolition is expected to be announced by the end of this year.

Meanwhile, legal challenges to the TMT continue to wend their way through Hawaii's courts. But the state's governor has said that the project has the authority to proceed, and unless the highest court rules otherwise, the TMT can begin construction again.

For now, the TMT and the demonstrators remain in an uneasy stand-off. Most recently, in mid-September, Mangauil and his colleagues agreed to leave their encampment partway up Mauna Kea, which they had occupied continuously since March. As *Nature* went to press, the TMT had not announced

when it might try to resume construction.

In the meantime, Simons continues to try building bridges whenever possible. In early August, he was one of more than 3,000 astronomers in Honolulu for a meeting of the International Astronomical Union. Most were not from Hawaii, and Simons arranged for Mangauil to give a private presentation to around 30 interested astronomers — "what I would call a one-hour classroom tutorial in his perspective", Simons says.

But Mangauil did more than that. A week before his presentation, he gathered some three dozen demonstrators outside the convention centre where the meeting was taking place. He and a UH historian held a press conference there, describing their grievances against the state, the university and the TMT in particular. After the talks were over and journalists had asked their questions, someone broke out a ukulele and led the crowd in traditional song.

About 20 metres away, some astronomers peered out curiously from inside the glass-walled conference centre. Most continued to travel up and down the main escalator on their way to poster sessions and talks about the Galactic Centre and the origin of the Universe.

One of the very few scientists who left the building to mingle with the demonstrators was Simons. He walked among them, head bowed in conversation. ■

Alexandra Witze writes for *Nature* from Boulder, Colorado.

COMMENT

SUPERCOMPUTING Approximate processing will advance modelling **p.32**

AUTUMN BOOKS Grande dame of food politics takes on the soft-drinks industry **p.34**



CLIMATE Geoengineering debated, with erudition and poetry **p.38**

EARTH Marking the centenary of the book that proposed continental drift **p.43**

TOP: BOTTOM RIGHT: COLD SPRING HARBOR LAB. LIBRARY & ARCHIVES; BOTTOM LEFT: ERIC GREEN



1989, Pre-HGP Banbury meeting. Francis Collins & James Watson in top row.



1990, Washington University School of Medicine. Madgnard Olsen & Eric Green.



1997, HGP meeting at Cold Spring Harbor. E. Green, R. Myers, J. Witkowski & R. Gibbs.

Twenty-five years of big biology

The Human Genome Project, which launched a quarter of a century ago this week, still holds lessons for the consortium-based science it ushered in, say **Eric D. Green, James D. Watson and Francis S. Collins.**

Twenty-five years ago, the newly created US National Center for Human Genome Research (now the National Human Genome Research Institute; NHGRI), which the three of us have each directed, joined forces with US and international partners to launch the Human Genome Project (HGP). What happened next represents one of the most historically significant scientific endeavours: a 13-year quest to sequence all three billion base pairs of the human genome.

Even just a few years ago, discussions surrounding the HGP focused mainly on what insights the project had brought or would bring to our understanding of human disease. Only now is it clear that, as well as dramatically accelerating biomedical research, the HGP initiated a new way of doing science.

As biology's first large-scale project, the HGP paved the way for numerous consortium-based research ventures. The NHGRI alone has been involved in launching more than 25 such projects since 2000. These have presented new challenges to biomedical research — demanding, for instance, that diverse groups from different countries and disciplines come together to share and analyse vast data sets.

It is easy for young researchers to forget that many of the problems they are trying to solve today had not even been thought about by their predecessors a quarter of a century ago. Equally easy to lose sight of are the insights that the HGP still offers to those pursuing big science projects. In fact, we think that the success of today's consortium-based science depends on six key lessons from the HGP.

SIX LESSONS

Embrace partnerships. By necessity, the HGP broke the mould of individual researchers toiling away in isolation to answer a small set of scientific questions. It also ran against the grain of hypothesis-driven research, focusing instead on the discovery of fundamental information that would inform many follow-on investigations.

The HGP brought together more than 2,000 researchers from many countries, disciplines and levels of seniority, with subgroups answering to different funding agencies. Success stemmed from: strong ▶

► leadership from the funders; the shared sense of the importance of the task; and the willingness of the researchers involved to cede individual achievements for the collective good¹.

Many consortium-based genomics projects followed. Among them are the 1000 Genomes Project, which is cataloguing sequence variants in the human genome (see pages 68 and 75), The Cancer Genome Atlas, which is characterizing the mutations responsible for cancer, and the Human Microbiome Project, which uses genome sequencing, among other techniques, to study microbial communities.

A frequent barrier to consortium-based science is the unwillingness of participants to embrace new partnerships. But various efforts — combined with the increasing realization that pooling data and resources can benefit everyone — are dismantling old norms.

Until recently, for instance, African genetics and genomics researchers collaborated most often with US or European scientists, and seemed less inclined to partner with other African researchers. A key objective of the Human Heredity and Health in Africa (H3Africa) initiative², which aims to enhance genomics research in Africa, has been to foster collaborations within Africa. The initial set of grants awarded by the US National Institutes of Health (NIH) and Britain's Wellcome Trust for the project in 2012 and 2013 established 29 collaborations involving 24 African countries; those numbers have since increased. H3ABioNet, a bioinformatics network that aims to facilitate the sharing of expertise, infrastructure and tools for analysing data across Africa, now involves 32 research groups in 15 countries.

Maximize data sharing. The HGP changed the norms around data sharing in biomedical research. Once large amounts of genome mapping and sequence data began to be generated, momentum quickly grew for establishing policies that shortened the time between the generation and release of data. These efforts culminated in adoption of the Bermuda Principles in 1996, when the heads of the major groups involved in the project agreed to submit genome-sequence assemblies above a certain size to a public database within 24 hours of generating them.

Such efforts have been built on in the years since. The principles were extended by the Fort Lauderdale Agreement in 2003. And in 2008, the NIH expanded its data-sharing expectations to include genome-wide association studies — analyses of common genomic variants in hundreds or thousands of people conducted to reveal variants associated with some trait of interest. In 2014, it started implementing an expansive Genomic Data Sharing Policy, which requires that almost all



Early days: a DNA-sequencing lab in 1994.

large-scale genomic data generated or analysed using NIH funds are shared.

Widespread sharing of data is throwing up new challenges. These include the computational and logistical difficulties of analysing and moving vast data sets; and in the case of human data (especially genomic and clinical), the problem of how to protect the privacy of research participants. Various initiatives are being pursued to address these problems.

The need for robust and powerful computing platforms is leading to rapid growth in the use of cloud computing in biomedical research, for instance. New resources are being proposed, such as a 'data commons' to house published and unpublished data³. And the Global Alliance for Genomics and Health, an international coalition established in 2013, is preparing an international Framework for Responsible Sharing of Genomic and Health-related Data⁴. This will take into account legal, ethical and technical considerations.

Plan for data analysis. Planning for the HGP had its flaws. In retrospect, one area that received insufficient attention early on was data analysis. The first

human genome sequence was produced in a piecemeal fashion. And to generate a contiguous sequence for each chromosome, thousands of individually assembled sequence segments (each around 100–300 kilobases) had to be stitched together computationally. The need for such a computational process (which turned out to be technically challenging) became apparent relatively late in the project. Through the heroic efforts of a small group of bioinformaticians, this task was accomplished in a matter of months. More care in planning would have made the endeavour much less stressful.

In recent years, several genomics projects (such as the 1000 Genomes Project and The Cancer Genome Atlas) have demonstrated how the early design of plans for data analysis can inform strategies for data generation. More recently, planning for the US Precision Medicine Initiative⁵ included considerable discussion about how best to merge and analyse the anticipated myriad data types — from electronic health records and genomic analyses to information from environmental monitors and wearable body sensors.

Prioritize technology development. In October 1990, the HGP participants pressed ahead, fully aware that the tools and methods for mapping and sequencing the human genome would need

“Waiting for absolute clarity about how the ultimate goals will be achieved risks missing opportunities.”

LEFT: HANK MORGAN/SPL; RIGHT: LAWRENCE BERKELEY NAT'L LAB/SPL



By 2006, DNA sequencing required much less manpower.

to be developed as part of the larger programme. In fact, the project catalysed the development of numerous crucial genomic technologies, and led to substantial innovations in molecular biology, chemistry, physics, robotics and computation, as well as to strategies for using tools and methods in innovative ways. In some cases, multiple incremental improvements were cobbled together to yield revolutionary advances, such as the capillary-based DNA sequencing instruments that were ultimately used to generate the first human genome sequence.

The need to foster technical innovations from the start is similarly crucial for today's large-scale projects. One effort leading the way in this respect is the US Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative⁶. With the overarching goal of revolutionizing our understanding of the human brain, the programme will focus initially on developing a new generation of tools for defining all the cell types in the brain, building maps of their connections, and recording signals from circuits that can be correlated with functions and behaviours.

Address the societal implications of advances. The founders of the HGP recognized that the information gained from mapping and sequencing the human genome could have profound implications

for society. The HGP thus became the first large-scale research project to include a component dedicated to examining broader societal issues, such as how to protect people's privacy and prevent discrimination. This arm of the project — known as ELSI (ethical, legal and social implications) research — was supported by about 5% of the NIH budget for the HGP⁷. It was the largest ever investment in bioethics research.

Societal and ethical considerations attend many of today's cutting-edge pursuits. High-profile examples include the use of the CRISPR/Cas9 gene-editing tool to alter the genomes of humans and other species, and the fast-tracking of clinical trial design for the rapid study of potential treatments during infectious outbreaks. Unfortunately, most consortium-based projects do not include a dedicated bioethics research programme as the HGP did. We think that as new large initiatives are launched, such programmes should be a key component.

Be audacious yet flexible. The goals of the HGP were bold. Given the lack of clarity on how exactly the human genome would be mapped and eventually sequenced, it was not surprising that the effort was viewed with some scepticism.

We believe that key to the HGP's success was the continued open-mindedness of the scientific leaders, and the

regular pauses they took to take stock. The initial five-year plan for the HGP was updated with revised plans in 1993 and in 1998. Individual HGP elements were regularly refined⁸.

Large projects with daring goals can prosper as long as overall objectives are grounded in explicit milestones, quality metrics and assessments. They also need a willingness to iterate plans as needed. Waiting for absolute clarity about how the ultimate goals will be achieved risks missing opportunities that present themselves only after researchers start work. This formula has become the norm for several large-scale projects, among them the BRAIN Initiative and the Precision Medicine Initiative.

GAME CHANGER

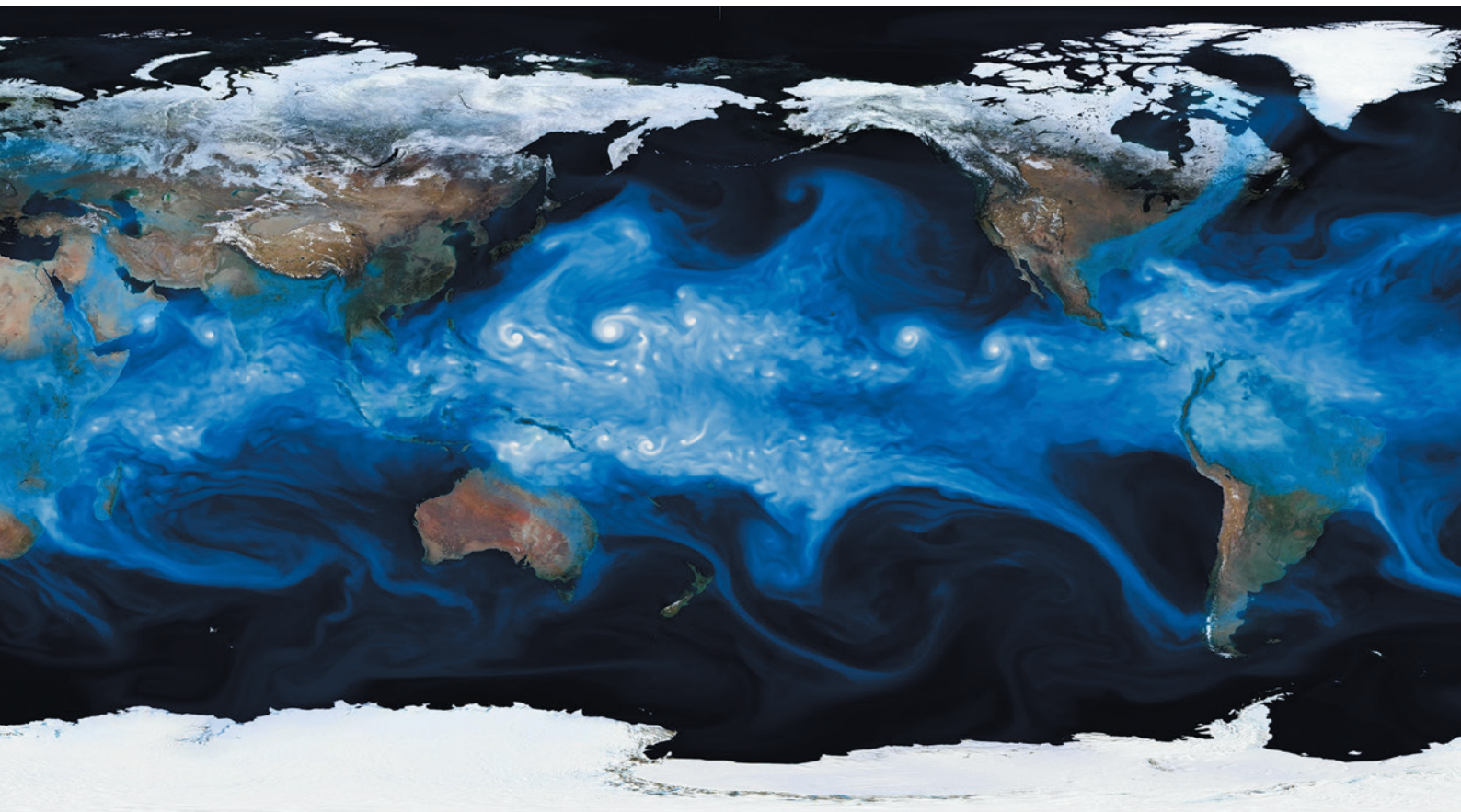
In the early 1990s — whether it was while leading the NIH's effort in the HGP (J.D.W. and F.S.C.) or working on the front line of the project (E.D.G.) — none of us foresaw that a major legacy of the HGP would be a new way of doing science.

During their careers, today's graduate students will probably witness and facilitate the unravelling of the molecular mechanisms for thousands of diseases, a revolution in cancer diagnosis and treatment, the maturing of microbiome science, the routine use of stem-cell therapies, and other spectacular biomedical advances.

The story of the HGP provides a valuable reminder that some of these advances will almost certainly trigger fundamental changes in the way that research is done — as well as a reminder of the importance of accepting and celebrating those changes. ■

Eric D. Green is director of the US National Human Genome Research Institute at the US National Institutes of Health, Bethesda, Maryland, USA. **James D. Watson** is chancellor emeritus at the Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA, and former director of the US National Center for Human Genome Research. **Francis S. Collins** is director of the US National Institutes of Health, Bethesda, Maryland, USA, and former director of the US National Human Genome Research Institute.
e-mails: egreen@nhgri.nih.gov; collinsf@mail.nih.gov

- Collins, F. S. *et al.* *Science* **300**, 286–290 (2003).
- H3Africa Consortium. *Science* **344**, 1346–1348 (2014).
- Stein, L. D. *et al.* *Nature* **523**, 149–150 (2015).
- Knoppers, B. M. *HUGO J.* **8**, 3 (2014).
- Collins, F. S. & Varmus, H. N. *Engl. J. Med.* **372**, 293–295 (2015).
- Insel, T. R. *et al.* *Science* **340**, 687–688 (2013).
- McEwen, J. E. *et al.* *Annu. Rev. Genomics Hum. Genet.* **15**, 481–505 (2014).
- Green, E. D. in *The Metabolic and Molecular Bases of Inherited Disease* 8th Edn (eds Scriver, C. R. *et al.*) 259–298 (McGraw-Hill, 2001).



A simulation of Earth's atmosphere generated by the Community Atmosphere Model.

Build imprecise supercomputers

Energy-optimized hybrid computers with a range of processor accuracies will advance modelling in fields from climate change to neuroscience, says **Tim Palmer**.

Today's supercomputers lack the power to model accurately many aspects of the real world, from the impact of cloud systems on Earth's climate to the processing ability of the human brain. Rather than wait decades for sufficiently powerful supercomputers — with their potentially unsustainable energy demands — it is time for researchers to reconsider the basic concept of the computer. We must move beyond the idea of a computer as a fast but otherwise traditional 'Turing machine', churning through calculations bit by bit in a sequential, precise and reproducible manner.

In particular, we should question whether all scientific computations need to be performed deterministically — that is, always producing the same output given the same

input — and with the same high level of precision. I argue that for many applications they do not.

Energy-efficient hybrid supercomputers with a range of processor accuracies need to be developed. These would combine conventional energy-intensive processors with low-energy, non-deterministic processors, able to analyse data at variable levels of precision. The demand for such machines could be substantial, across diverse sectors of the scientific community.

MORE WITH LESS

Take climate change, for example. Estimates of Earth's future climate are based on solving nonlinear (partial differential) equations for fluid flow in the atmosphere and oceans. Current climate simulators — typically with

grid cells of 100 kilometres in width — can resolve the large, low-pressure weather systems typical of mid-latitudes, but not individual clouds. Yet modelling cloud systems accurately is crucial for reliable estimates of the impact of anthropogenic emissions on global temperature¹.

The resolution of this computational grid is determined by the available computing power. Current petaflop computers can perform up to 10^{15} additions or multiplications — floating-point operations — per second (flops). By the early 2020s, next-generation exaflop supercomputers, capable of 10^{18} operations per second, will be able to resolve the largest and most vigorous types of thunderstorm². But cloud physics on scales smaller than a grid cell will still have to be approximated, or

LAWRENCE BERKELEY NATL LAB/DATA: MICHAEL WEHNER (LEND)/VISUALIZATION: PRABHAT (LEND)

parametrized, using simplified equations.

Errors introduced by such parametrizations proliferate and infect calculations on larger scales. In climate simulators, these errors can be represented by introducing stochastic noise into the computational equations³. Hence, climate prediction is inherently probabilistic.

The main obstacle to building a commercially viable 'exascale' computer is not the flop rate itself but the ability to achieve this rate without excessive power consumption. Early estimates suggested that such computers would consume about 100 megawatts — the output of a small power station. A key challenge in recent years has been to make exascale computers more energy efficient.

Energy is required to perform basic arithmetic operations in computers. As microprocessors shrink to the nanometre scale, extra energy is needed to overcome thermal noise and even cosmic-ray strikes. By turning down the voltage, processors can be switched from deterministic to probabilistic calculators. For example, based on contemporary chip technology, a fourfold reduction in power can result in less than a 1% chance that a computational step will be incorrect⁴.

More importantly, energy is also used to move data from one part of a computer to another. The amount needed to accomplish this is proportional to the number of bits used to represent individual pieces of data. The 'gold standard' for variables taking real-number values is the 64-bit double-precision representation. Although supercomputers used in scientific computation also support 32-bit representations, they do not support representations with less than 32 bits, presumably because vendors perceive there to be little demand for such variable types.

There is no point, however, in being more deterministic or precise than is justified by the overall accuracy of the computational code; in the case of climate models, this accuracy is limited by errors from the parametrization schemes. Although 64-bit precision may be appropriate in representing variables associated with planetary-scale jet streams or weather systems that are hundreds of kilometres across, it is a waste of computing and energy resources to use this precision to represent smaller-scale circulations approaching the resolution limit of a climate model⁵. This is important because, collectively, the small-scale computations and sub-grid parametrization formulae dominate the cost of a climate simulation.

The energy liberated by not performing overly exact calculations could be put to more productive use. This requires a new type of supercomputer. Like current ones, such a machine would be massively

parallel, comprising many millions of individual processing units. A fraction of these would enable conventional, energy-intensive and deterministic high-precision computation. Unlike conventional computers, the remaining processors would be designed to take on low-energy probabilistic computation with lower-precision arithmetic, the degree of imprecision and inexactness being variable.

By using power more efficiently, a hybrid exascale supercomputer could extend the dynamic range of climate models to below the kilometre scale, allowing deep convective clouds to be well resolved. This would enable more reliable probabilistic predictions of Earth's future climate.

WIDER BENEFITS

Inexact hybrid computing has the potential to aid modelling of any complex nonlinear multiscale system — from galactic and stellar evolution to tokamak plasmas and combustion in jet engines.

Living systems may already be aware of the benefits. The brain achieves its prodigious signal-processing capabilities using around 20 watts, less power than a typical light bulb. Axons — nerve fibres — and the ion channels that amplify the electrical signals that pass along them are of molecular dimensions and require little energy. The diameter of a typical human axon, 0.1 micrometres, is so small that the signals it contains are susceptible to thermal noise and hence to random fluctuations⁶. Although such noise is often considered to hinder the operation of the nervous system, in some circumstances it might offer an advantage⁷.

When considering computational systems that might mimic the brain, mathematician Alan Turing suggested that it would be "wise to include a random element in a learning machine" and provided a simple theoretical example to back up his claim⁸. It is now well known that adding noise can make algorithms more efficient⁹. For example, in the classic travelling salesperson problem — calculating the shortest possible route through multiple cities — adding a random noise component can reduce the overall computation time needed to find a solution⁹.

These considerations suggest that energy-efficient hybrid computing is indispensable for modelling the brain, and hence for understanding cognition. Indeed, it has been suggested that human creativity arises from a close synergy between low-energy computation (in which the operation of the brain is

susceptible to the brain's own thermal noise) and higher-energy determinism (in which the implications of partially random cognitive jumps can be explored algorithmically in localized parts of the brain)¹⁰. If this is the case, then human creativity might be a by-product of evolutionary pressures to optimize the use of energy available to power the brain.

PARADIGM SHIFT

Supercomputer manufacturers are driven by commercial pressures. The type of computer architecture I envisage will be built only if there is sufficient demand from the scientific community.

Scientists who use the top tier of supercomputers across a range of disciplines need to assess the extent to which the conventional — energy-intensive — deterministic approach to computation is becoming a bottleneck. To do this, they need to quantify the impact of inherent uncertainties associated with approximations in their computational codes. They can then estimate how much the levels of computational inexactness and indeterminism can be increased before the impact of these factors exceeds that of the inherent uncertainties.

Computer vendors should begin by marketing processors and scientific computing libraries that make efficient use of mixed-precision representations of real-number variables. Joint funding from research councils and the private sector should be made available to catalyse these developments.

High-performance computation is rapidly overtaking traditional experimentation in many scientific disciplines. In designing the next generation of supercomputers, we must embrace inexactness if that allows a more efficient use of energy and thereby increases the accuracy and reliability of our simulations. ■

Tim Palmer is a Royal Society research professor of climate physics and co-director of the Oxford Martin Programme on Modelling and Predicting Climate at the University of Oxford, UK.
e-mail: tim.palmer@physics.ox.ac.uk

1. Mauritsen, T. & Stevens, B. *Nature Geosci.* **8**, 346–351 (2015).
2. Palmer, T. *Nature* **515**, 338–339 (2014).
3. Palmer, T. N. Q. J. R. Meteorol. Soc. **138**, 841–861 (2012).
4. Palem, K. V. *Phil. Trans. R. Soc. A* **372**, 20130281 (2014).
5. Düben, P. D. & Palmer, T. N. *Mon. Weath. Rev.* **142**, 3809–3829 (2014).
6. Faisal, A. A. & Laughlin, S. B. *PLoS Comput. Biol.* **3**, e79 (2007).
7. McDonnell, M. D. & Ward, L. M. *Nature Rev. Neurosci.* **12**, 415–426 (2011).
8. Turing, A. M. *Mind* **236**, 433–460 (1950).
9. Hoos, H. H. & Stützle, T. *Stochastic Local Search: Foundations and Applications* (Elsevier, 2005).
10. Palmer, T. N. & O'Shea, M. *Front. Comput. Neurosci.* **9**, 124 (2015).

AUTUMN BOOKS



NUTRITION

Dominions of fizz

David Katz applauds an analysis of the carbonated-drinks industry and public health.

If any one name evokes unfettered truths about the sociopolitical machinations of 'Big Food', it is that of Marion Nestle, professor of nutrition, food studies and public health at New York University. Author of *Food Politics* (Univ. California Press, 2002) and the blog of the same name, she held senior positions in US food policy in the 1980s and 1990s, sitting, for example, on the 1995 US Dietary Guidelines Advisory Committee. Her writing exerts a powerful influence on almost all other contributors in this realm.

Nestle's latest, *Soda Politics*, addresses carbonated, non-alcoholic, sweetened beverages as an emblem of modern wars focused on food, politics, policy, personal choice and culture. This concentrated source of sucrose, high-fructose corn syrup and

calories, free of any nutritional attributes, accounts for one-third of all US sugar intake. *Soda Politics* is what those who know Nestle and her work would expect. It is thorough and thoughtful, careful and comprehensive, exacting and erudite — and only rarely surprising. She elaborates opposing views before rendering her generally moderate verdicts, such as: "Sugar is neither a poison nor entirely harmless."

After defining her terms, Nestle distils what is agreed and what is contentious regarding the health effects of soft drinks, and provides an overview of the industry (valued at anything from US\$200 billion to \$800 billion globally) and its characteristic responses to public health. She covers the scientific evidence on health effects, the industry's impact

on the environment and the preferential marketing of soft drinks to children, specific ethnic groups and poor people, for instance at sporting and cultural events — strategies that Nestle characterizes as "softball".

A prominent theme in *Soda Politics* is the correspondence between the tactics of the soft-drinks and tobacco industries. Both use "hardball" strategies such as litigation, lobbying of Congress, and front groups such as New Yorkers Against Unfair Taxes, established by the beverage industry to oppose a soft-drinks levy. Nestle asserts that these interests "forge alliances with health organizations and researchers to make the science appear confusing and to silence criticism" — tactics that stretch back to the 1970s and beyond. She cites the work of beverage-industry-funded

fanfare, Nestle relates, “rescued sports drinks, sugar-sweetened waters, and the machines that sold them”, while helping soft-drink companies to sidestep a class-action lawsuit. As Nestle shows, this lawsuit was abandoned with Clinton’s encouragement when the beverage industry agreed to the terms brokered by the Alliance.

We also hear of close ties between the leadership of the Robert Wood Johnson Foundation, a philanthropic body focusing on many aspects of health and health care, and that of PepsiCo. Nestle writes that the foundation’s president and PepsiCo’s chief executive routinely sit together at public events. We are told that as much as \$4 billion in food stamps under the Supplemental Nutrition Assistance Program is spent each year on soft drinks. And we learn of the unexpected alliance of entities that oppose remedying this with policy, such as food retailers that profit from the programme. We get a bracing reality check regarding front groups. For instance, the non-profit education and advocacy organization the American Council on Science and Health, Nestle tells us, “depends heavily on funding from corporations that have a financial stake in the scientific debate it aims to shape”. Coca-Cola is a significant sponsor.

Nestle’s decisive opinions slice through a number of polarized controversies in public health. She asserts that “it is so well established that sodas and other sugary drinks contribute to higher calorie intake, weight gain, obesity, and type-2 diabetes that stopping drinking them is the first line of defense against any of these conditions”. Amen. She states that diet drinks, which now account for 30% of US soft-drink sales, have not been shown to help most people to control their weight. And she points out that although high-fructose corn syrup and sucrose differ in how they deliver free fructose and in their specific metabolic effects, it is unclear that the differences matter much to public health in light of the overall excess. My own work in this area has led me to the same conclusion.

Occasionally, *Soda Politics* serves up genuine surprises. Coca-Cola, for instance, markets 3,500 products under 500 brand names in more than 200 countries. Yet despite fierce brand loyalties, the products of the leading manufacturers are consistently

indistinguishable in blind taste tests. As a result of industry obfuscation, which keeps relevant data proprietary and shielded from public view, researchers are not sure how much the modern citizen drinks, only that it is a lot (average per capita US intake has been estimated at nearly 170 litres per year). Each 350-millilitre portion contains 10 teaspoons of sugar — other ingredients may serve principally to mask this

extreme sweetness. Nestle also briefly discusses 4-methylimidazole, a by-product of the caramel colouring used in some soft drinks, which has been deemed a potential carcinogen by the US National Toxicology Program following thus-far equivocal findings in a two-year rodent study. The US Food and Drug Administration is currently reviewing the range of data available on the compound.

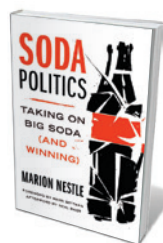
For me, the single most stunning and appalling revelation

comes in the section about environmental impact and industry responses to it. It is that between 340 and 620 litres of water are used for every litre of soft drink produced, about 20% of that related to packaging. Despite such disturbing revelations, *Soda Politics* is not discouraging. The parallels between the practices of the soft-drink and the tobacco industries can inform strategies for winning this public-health battle, pointing to moves such as banning television advertising. Throughout the book, Nestle provides tactics for practical, local advocacy, such as working with school wellness committees and engaging local policymakers. And since 2002, the proportion of US citizens who say that they avoid soft drinks has risen by 20%, reaching nearly two-thirds of the population.

Nestle cannot attribute that trend to any one action; it is the aggregate effect of many, and of increasing awareness. The soft-drink industry is, however, vast and shrewd, profitable, pervasive and powerful. For public health to prevail over soda politics as usual, we have miles to go. This book is the richly drawn map of how to get there, from here. ■

David Katz is the founding director of the Yale-Griffin Prevention Research Center in Derby, Connecticut, founder of the True Health Coalition (<http://glimmerinitiative.org/>) and president of the American College of Lifestyle Medicine. His latest book is *Disease Proof*. e-mail: davkatz7@gmail.com

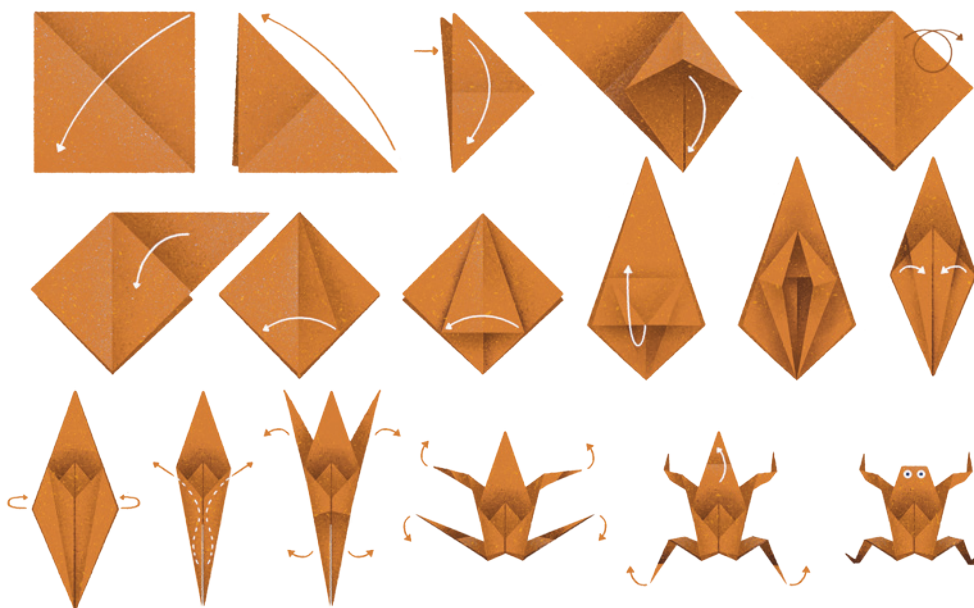
**BETWEEN
340 AND 620
LITRES
OF WATER ARE USED FOR
EVERY LITRE
OF SOFT DRINK
PRODUCED.**



**Soda Politics:
Taking on Big Soda
(and Winning)**
MARION NESTLE
Oxford University
Press: 2015.

researchers who examine the effects of soft-drink consumption on health, and highlights how their results consistently diverge from findings of studies with unconflicted funders. She backs up every argument abundantly; this is a hefty, well-researched book.

Nestle’s blunt assessment of current interactions between the soft-drink industry and certain luminaries of public health and public policy is provocative. She tells us about the Alliance for a Healthier Generation — founded jointly by the American Heart Association and the Clinton Foundation (a non-profit group set up by former US president Bill Clinton to help people meet “challenges of global interdependence”). It was, she writes, formed to negotiate policies on selling soft drinks in schools with the beverage industry. The deal reached, and announced with



EVOLUTION

Parsing the cycles of change

Mark Buchanan examines a treatise on evolution as central to processes in a vast, varied range of domains.

Evolution is an almost magical idea. First proposed by Charles Darwin in 1859 as an explanation for the manifold diversity of biology, the concept has turned out to be much more profound than its inventor could have imagined. Evolution is a general strategy, or class of strategies, for finding solutions to very difficult problems through iterative, combinatorial exploration in high-dimensional spaces of possibilities. Organisms evolve, and so do algorithms for image recognition or for financial trading.

Matt Ridley, an accomplished science writer and Conservative member of the UK

House of Lords, has explored the power of evolution in biology in half a dozen books. In his latest, *The Evolution of Everything*, Ridley makes a powerful argument that evolution in a more general sense has created most of the things that we treasure — from modern technology to decent government and reasonably stable economies. He also ponders the mystery of why, despite this overwhelming evidence for the value of evolution in design, so many people still long for the apparent order of top-down planning and control, solutions designed and implemented by policy experts.

Over 16 chapters, Ridley explores



The Evolution of Everything: How New Ideas Emerge

MATT RIDLEY
Fourth Estate: 2015.

processes that involve incremental change through trial and error. He considers the evolution of the Universe, morality, the economy, technology, money and more — even the future. In each, he examines how attempts to solve human problems through logical planning and purposeful intervention so often fail.

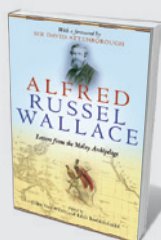
Take overpopulation. In the 1960s and 1970s, a number of writers — most prominently, the ecologist and demographer Paul R. Ehrlich — proclaimed that global famines would soon devastate humanity unless drastic action were taken to stop the population explosion. The problem, one expert suggested, required the creation of a planetary authority with responsibility “for determining the optimum population for the world and for each region”. The idea was a non-starter, and even trying to implement it would probably have caused immense suffering. As Ridley argues, it was the evolutionary inventiveness of science and changing human practices that offered a solution, at least temporarily. We found much more efficient agricultural methods, and people, as they grew more prosperous, started to have fewer children.

In this case, and in many others that Ridley examines, solutions to important human issues were discovered not through conscious planning, but through undirected experimentation. We defeated the dark of night through the slow accumulation of many discoveries — fire, the production of metals, the steam engine, vacuum technology and so on — none of which were expressly aimed at illumination. Similarly, nearly all human societies have created powerful, flexible written languages for communication — not by design, but through slow adaptation, adjustment and modification.

Ridley is generally correct. The world is teeming with systems — anything from the Internet to New York City traffic — that

**NEW IN
PAPERBACK**

*Highlights of this
season's releases*



Alfred Russel Wallace: Letters from the Malay Archipelago

Edited by John van Whye and Kees Rookmaaker (Oxford Univ. Press, 2015)

Alfred Russel Wallace led an adventurous life in science, from insect spotting in a Borneo swamp to exploring Ternate island, Indonesia, where he independently developed a theory of natural selection. This collection of correspondence from 1854 to 1862 covers his fateful travels. The letters (which took six weeks to arrive), to and from Wallace's family and Charles Darwin, shed light on the controversy over precedence of the theory, as well as the malaria and other hardships that Wallace suffered for his work.

are much too complex to engineer and control with top-down thinking. And his book offers revealing examples of how evolution has improved approaches across essentially all fields, from software design and telecommunications to the economics of housing and basic human morality. *The Evolution of Everything* will be enjoyed by anyone interested in the origins of order and organization in human societies, and how we might put evolutionary forces to better use in managing our lives and communities.

One thing that I liked less about the book, however, is how Ridley's political views often intrude on his arguments. His examination frequently gives way to complaints about all manner of things that he — a libertarian — despises. Too much government and meddling in health care; too many taxes and layers of social policy to protect people. Ridley manages to blame the good intentions of left-leaning people for the persistence of global poverty, for the demise of the British health-care system, even for fascism. Most of the intelligent public, Ridley grouches, believes that government is the foundation of all that is good, and is generally infallible.

Does anyone actually believe this? Most people just think that government does some necessary and useful things — helping to ensure the stability of the financial system, for example, and providing basic levels of education. Most economists think the same. This aspect of the book will no doubt appeal to the libertarian element in right-wing organizations, but for many readers, the asides will interfere with the discussion.

If you filter out the political cheerleading, Ridley's argument emerges as edifying. It is almost certainly true that solutions to our most pressing problems — from global poverty to climate change — are not going to spring from the mind of any lone genius or planning committee. We will find them through the collective tinkering and evolutionary exploration of tens of millions of diverse minds working together. ■

Mark Buchanan is a science writer based in Abbotsbury, UK. His latest book is *Forecast*. e-mail: buchanan.mark@gmail.com

PHYSICS

Two shades of physics

Robert P. Crease contrasts a physicist's account of awe with a historian's reality check.

These two concise tours of physics are delightful, each in their own way. In *Seven Brief Lessons on Physics*, physicist Carlo Rovelli appreciates the field's beauties in an expansion of articles he wrote for the Italian newspaper *Sole 24 Ore*. Science historian John Heilbron's *Physics* surveys the discipline from ancient times to today.

Rovelli begins by relaying his excitement at discovering the general theory of relativity for the first time, in the gnawed pages of a textbook he had used to plug mouse holes. Reading it on a beach in Italy, he was inspired by its disclosure of a simpler, deeper order to the Universe — the gravitational field is not diffused through space, but is space. It was “as if a friend was whispering into my ear an extraordinary hidden truth”.

He writes evocatively of the theory's many wonders: exploding universes, space collapsing into bottomless holes, time sagging and slowing and the unbounded extensions of interstellar space rippling and swaying “like the surface of the sea”. We are immersed not in an invisible rigid infrastructure, but in “a gigantic flexible snail-shell”. The metaphors are vivid, the visions dramatic. When this book was first published in November 2014 in Italy, it outsold E. L. James's blockbuster novel *Fifty Shades of Grey* (Vintage, 2011).

Through chapters on quantum principles, cosmology, particles, quantum gravity and thermodynamics, Rovelli maintains the awestruck tone of a practising physicist. Only in a final chapter on the place

Seven Brief Lessons on Physics

CARLO ROVELLI
Allen Lane: 2015.

Physics: A Short History from Quintessence to Quarks

JOHN L. HEILBRON
Oxford University Press: 2015.

of humans “in this great fresco” does this stance lead him astray. It makes it hard to explain why some people struggle to comprehend science, and even distrust it. It tempts him into scientism — regarding the world that science describes as the real

one. The flow of time, he suggests, is “absent from descriptions of the world”. Yet philosophical ‘lived time’ — the process of anticipating the future out of a past to allow the human experience of the present — is a fundamental condition of being human. It allows us, among other things, to create and marvel at scientific frescos.

Placing himself as observer rather than participant, Rovelli forgets where he stands.

Heilbron's *Physics* is different in topic and tone. He uses the Greek word *physis* to name the ancient field, then traces how it morphed into physics. *Physis* seamlessly folded in astronomy, psychology and zoology; its idea of cause included form, purpose and the stuff of which things were made, as well as pushes and pulls. From this, Aristotle developed a ‘theory of everything’, which explained almost all phenomena experienced by humans, from the growth and behaviour of plants and animals to the patterns made by heavenly bodies. It included a deity that drew things into motion; ▶

THE FLOW OF TIME IS
ABSENT
FROM DESCRIPTIONS
OF
THE WORLD.



The Quantum Moment

Robert P. Crease and Alfred Scharff Goldhaber
(W. W. Norton, 2015)
Philosopher Robert Crease and physicist Alfred Goldhaber reveal how quantum theory has pervaded popular culture, from quantum poetics to television's *Quantum Leap* (see Jim Baggott's review: *Nature* **513**, 308–309; 2014).



Adventures in the Anthropocene

Gaia Vince (Milkweed, 2015)
The human epoch is in full swing, with a population of 8 billion looming. In search of sustainability, journalist Gaia Vince travelled to six continents and found much to foster hope — such as the Ugandan farmer who feeds livestock on a by-product of her sunflower crop.

► and ‘quintessence’, a fifth element (in addition to the familiar earth, air, fire and water), which was needed to keep the theory consistent, explaining, for example, why heavenly bodies move in circles rather than in straight lines.

Physica did not become physics simply as a result of observant people adding pieces to a puzzle. It required transformations in the social ecosystem, such as who pays for knowledge and why; its social applications; and how it is communicated. *Physica* got a big boost from the Islamic world, where Aristotle’s concept was highly regarded and translated into Arabic around the ninth century. But physics began to acquire its eventual outline in the West after the sixteenth century, with the generation of Francis Bacon, Galileo and René Descartes.

Fostered by the needs of centralized, bureaucratic states, the discovery of new worlds, the spread of universities and new industrial applications, the emergence of physics as we know it today was a process of “dedefying and deanthropomorphizing nature”. Now, God is marginalized and ‘dark energy’, our new quintessence, is needed to make sense of it all. A theory of everything is an ever more remote goal.

Heilbron does not sneer at *physica*, but carefully examines it and the ecosystem in which it thrived. By the book’s end, physics has split off into so many branches — radar, Earth science, space probes, accelerators, meteorology and so on — permeating so many spheres of human life that we begin to lose sight of the field as something coherent. And that is the point.

Whereas Rovelli’s feel-good book ends with us gazing in wonder at the edge of “the ocean of the unknown”, Heilbron leaves us rooted in lived reality. “Physics has given civilization a somber, disturbing, and challenging world picture, many fertile and some terrifying inventions, and notice of responsibility for the outcome of the human story.” If it, too, outsells *Fifty Shades*, there is hope for humanity yet. ■

Robert P. Crease is a professor in the Department of Philosophy at Stony Brook University, New York.
e-mail: robert.crease@stonybrook.edu

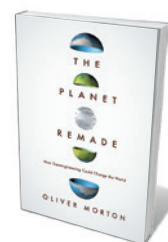
GEOENGINEERING

Journey into geopoetry

Jane C. S. Long relishes an erudite exploration of the people and principles of climate intervention.

Several authors have tackled geoengineering — the idea of harnessing science and technology to cool our overheated planet. In the 2010 *How to Cool the Planet* (Houghton Mifflin Harcourt), Jeff Goodell told the personal stories of geoengineers. One, physicist David Keith, described how his interest in climate modification is grounded in a desire to preserve nature in *A Case for Climate Engineering* (MIT Press, 2013). Jack Stilgoe discussed responsible governance of geoengineering in *Experiment Earth* (Routledge, 2015). Eli Kintisch covered the history of potential solutions and their developers in *Hack the Planet* (Wiley, 2010). Stewart Brand described intervention as inevitable in *Whole Earth Discipline* (Atlantic, 2010), stating: “We are as gods and we might as well get good at it.”

But if you are going to read one book on climate engineering, it should be *The Planet Remade*. Oliver Morton, briefings editor at *The Economist*, starts by asking: do you think climate change is a problem, and the energy system easy to change? Using this dialectic, he explores the thesis that the climate crisis cannot be solved, but could be managed. There follows a journey through the people and principles of climate science and intervention, the natural history of carbon dioxide, engineering of the nitrogen cycle and the backstory of weather modification. Morton speculates about the ethical, political and social implications if climate intervention became available. The book finishes with a range of scenarios — including one that could end well for Earth and a frank discussion of what could go wrong. *The Planet Remade* is as much an exploration of science and



The Planet Remade: How Geoengineering Could Change the World

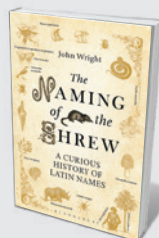
OLIVER MORTON
Granta: 2015.

engineering as it is of people and attitudes.

Most climate engineering proposes to change the radiation balance of Earth so that less radiation gets in, or more escapes. Techniques include spraying reflective aerosols into the stratosphere, brightening clouds with salt spray or sequestering greenhouse gases.

Morton traces the idea to the cold war, when scientists including physicist Edward Teller began to fear that a nuclear war would cause a hellish conflagration, darkening the skies and creating a ‘nuclear winter’. Efforts to understand this led to the birth of modern climate science — which in turn supported 1980s warnings about climate change by the likes of atmospheric physicists James Hansen and Stephen Schneider. Given clear evidence that volcanic eruptions can send enough reflecting sulfur particles into the stratosphere to cool Earth noticeably, it was not a great leap for some scientists to contemplate intentionally using sulfur to counteract greenhouse-gas emissions. *The Planet Remade* encourages researching this idea and others to learn more about their effectiveness, feasibility and advisability.

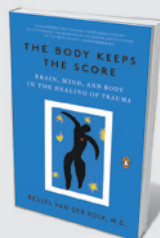
Climate engineering evokes very disparate and strong reactions. As Morton shows, some people, such as Keith, are keen to start intervention, whereas others, such as ethicist Clive



The Naming of the Shrew: A Curious History of Latin Names

John Wright (Bloomsbury, 2015)

Ba humbug! is not a curse but a snail, and bananas are a “taxonomic nightmare”. Fungus fanatic John Wright digs into taxonomy’s origins, including Carl Linnaeus’s overtly sexual plant-ordering system, based on reproductive parts.



The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma

Bessel van der Kolk (Viking, 2015)

Violence, abuse or conflict can burn trauma into memory. Psychiatrist Bessel van der Kolk reveals how severe stress rewires the brain, and suggests therapies from breathing techniques to eye-movement desensitization and reprocessing.



Hamilton, abhor the enterprise. Brand feels that the only option is to manage the global environment — which many fear would fail, because humanity lacks the wisdom and capability for the task. Morton weighs these perspectives with sympathy. He takes pains to find value in each, while maintaining his own insight gleaned from knowledge of the natural world, social thought, literature and science fiction, science and politics, scientific history and the scientists making that history. Lively anecdotes make clear that, as a journalist, Morton has known many of these people personally. Who else could tell us that the substance ice-nine in Kurt Vonnegut's novel *Cat's Cradle* (Holt, Rinehart & Wilson, 1963) was based on cloud-seeding research by the novelist's brother, Bernie?

His prose is sometimes hard to parse, but poetic — or “geopoetic”, as he would have it. In a moving passage, Morton explains that he has not said “we” because the world population has yet to unite to counterbalance climate change, although he hopes that it will. His hope is embodied in the beauty of elegant engineering and joy in a world of thriving life:

THE WORLD POPULATION HAS YET TO UNITE TO COUNTERBALANCE CLIMATE CHANGE.

“a reimagining of how humans and nature can intermingle, a new consciousness of what can be done for the planet rather than a blind deference to what are claimed to be its limits”. He sees the future as creating “a we... that can set a better course”. I share Morton's belief that contemplating climate intervention could help humanity to become a “we” that acts on the need to take responsibility for our planet, with or without geoengineering.

For a potentially harrowing topic, serendipity and fun abound. Plentiful and erudite footnotes are richly entertaining. To quote a favourite, in discussing controversy over defining the start of the human-influenced

Anthropocene epoch, Morton notes: “when scientific publications refer to an event happening 2,500 or 5,000 years ‘before present’ (BP), they actually mean before 1950. If 1950 were chosen as the beginning of the Anthropocene, then the Anthropocene would... be in a condition of permanent futurity, hanging unsupported in the air like a Wile E. Coyote that has run over the cliff at the end of history.”

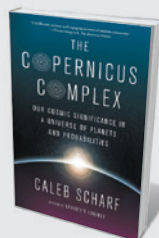
Who should read *The Planet Remade*? In some ways it is too technical for a lay audience, but too important to be reserved for experts. Anyone with a knowledge of the climate problem would benefit; it also works as a primer on energy, climate science and Earth-system science. I have a long list of people to whom I will be recommending it. ■

Jane C. S. Long works on reinvention of the energy system and geoengineering. She has retired as associate director for energy and environment at Lawrence Livermore National Laboratory in California, and was formerly dean of the Mackay School of Mines at the University of Nevada, Reno. e-mail: janeclong@gmail.com



From Eve to Evolution

Kimberly A. Hamlin (Univ. Chicago Press, 2015) Science historian Kimberly Hamlin shows how nineteenth-century US feminists used Darwinian evolutionary theory to argue for equality. Eliza Gamble, for example, put women's choice at the forefront of male–female attraction (see Sarah S. Richardson's review: *Nature* **509**, 424; 2014).



The Copernicus Complex

Caleb Scharf (Farrar, Straus and Giroux, 2015) Are we cosmically insignificant or the centre of the known Universe? Skipping from molecules to Moon landings, astrobiologist Caleb Scharf puts life on Earth under the microscope and concludes that humans are unique but unexceptional (see Mario Livio's review: *Nature* **512**, 368–369; 2014).



Origins: The scientific story of creation

JIM BAGGOTT

Oxford University Press: 2015.

Dark Matter and the Dinosaurs: The Astounding Interconnectedness of the Universe

LISA RANDALL

Ecco: 2015.

uses the analogy of knowing that a celebrity is near because of disrupted traffic and crowds of phone-wielding people. Her strong opinions — even ones I question, such as suggesting that “transparent matter” might be a better name than dark matter — liven the narrative.

The story begins with the Big Bang. What it is — the origin of matter, energy, space and time as Einstein’s general theory of relativity has it, the emergence of space and time as string theory might posit, or the outcome of a previous cycle of cosmic evolution — remains to be determined. Inflation follows: a burst of expansion that smooths and flattens the Universe and stretches quantum fluctuations to astrophysical size, to become the seeds for all structure in the Universe. The details have yet to be revealed. But evidence for inflation is growing, particularly in measurements of tiny variations in the temperature of the cosmic microwave background radiation.

This ‘quark soup’ phase lasts a microsecond, followed by nucleosynthesis and the formation of the lightest elements at 3 minutes. Atoms form at 380,000 years. Then gravity amplifies lumpiness in the distribution of matter to become galaxies, clusters of galaxies and superclusters, with the first stars and galaxies emerging at around 500 million years. The Sun forms some 9 billion years later.

Now the narratives turn to ‘local’ events: Solar System formation, Earth’s cooling, the emergence of oceans — and, 3.5 billion years ago, the first life forms. Important questions remain. Where did organic material originate? How did the transition from inorganic to organic occur? What was the last universal common ancestor, which Charles Darwin described as the primordial form from which all living things on Earth descend? From here, the pace quickens: multicellular organisms, atmospheric oxygenation around 2.5 billion years ago, sex as a mechanism for gene exchange, the emergence of primates shortly after the dinosaurs’ demise and,

COSMOLOGY

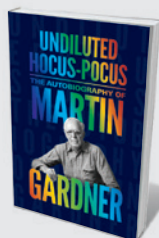
A story of cosmic proportions

Michael S. Turner weighs up two distinctive popular books on the evolution of the Universe.

Jim Baggott’s *Origins* and Lisa Randall’s *Dark Matter and the Dinosaurs* recount the greatest story ever told: the evolution of the Universe since the Big Bang. This rich cross-disciplinary tale reminds us that astronomy, physics, chemistry, geoscience, biology and neuroscience are interconnected. The books cover the same ground in very different styles. Baggott, a chemist turned science writer, takes the reader on a linear, 13.8-billion-year

journey. His textbook-like treatment abounds with excellent visuals, from charts to lithographs. At its best, *Origins* reminds me of Richard Holmes’s marvellous *The Age of Wonder* (HarperCollins, 2008).

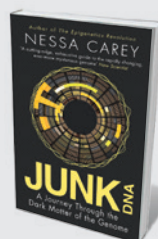
Randall, a particle physicist and cosmologist, makes the epic trip more succinct and conversational, interspersing her passions, perspectives and creative analogies. Describing how astronomers ‘see’ dark matter, she



Undiluted Hocus-Pocus

Martin Gardner (Princeton Univ. Press, 2015)

Zealously debunking science fads and declaring his bafflement at the human brain, maths writer Martin Gardner was on fine form in this posthumous memoir. As it reveals, his *Scientific American* column was just a piece of his life’s puzzle (see David Singmaster’s review: *Nature* **501**, 314–315; 2013).



Junk DNA

Nessa Carey (Icon, 2015)

If only 2% of human DNA is technically ‘useful’ in coding for proteins, what is the other 98% for? Geneticist Nessa Carey uses Jackson Pollock paintings and baseball bats to explain how ‘junk’ DNA keeps the body functioning (see Nathaniel Comfort’s review: *Nature* **520**, 615–616; 2015).

some 200,000 years ago, *Homo sapiens*.

Baggott ends at consciousness, that frontier of interdisciplinarity. But he fails to ask whether intelligent life is a convergent property of evolution. Given that evolution involves dominating local resources, the Universe may teem with 'dumb' life, while intelligent life remains exceedingly rare.

Earth's prehistory was marked by five major extinctions, identified in 1982 by palaeontologists David Raup and John Sepkoski. The Cretaceous–Palaeogene extinction 66 million years ago, which killed the dinosaurs, is the best known. Physicist Luis Alvarez and his geologist son Walter proposed that the cause was an asteroid impact, an idea met with scepticism until the mid-1990s, after a crater fitting the bill was identified in Mexico. Such an impact can alter conditions on Earth for tens of years, through a global dust cloud, firestorms and other after-effects: species ill-suited to such dramatic change go extinct.

Raup and Sepkoski also put forth evidence that extinction events occur roughly every 30 million years. This is now generally accepted, but there is no agreed mechanism. Randall and her collaborator Matt Reece offer

a hypothesis. They posit that there are two kinds of dark matter: the ordinary one, whose gravity binds galaxies and galaxy clusters, and a 'social' form that also interacts with its own kind. The social dark matter forms a thin disk of material in our Galaxy whose gravity can shake things loose in the outer depths of the Solar System when it crosses the Galactic disk, every 30 million years or so. Randall admits that the idea is a long shot, although testable. This aspect of *Dark Matter and the Dinosaurs* conveys the excitement and uncertainty of cutting-edge, big-idea research.

In a chapter called 'The cosmic imperative', Baggott implies that the evolution of life is an inevitable consequence of chemistry, despite our not knowing precisely how it occurred. This reminded me of physicist Murray Gell-Mann's dictum "Everything not forbidden is compulsory" (borrowed from novelist T. H. White), which describes the importance of symmetry principles in particle physics: they set the basic rules, but not the detailed outcomes. A rich set of rules (think chess) can lead to complex and interesting outcomes. I would take this further: the Universe is governed by physical laws that permit a rich set of behaviours, resulting

in its inevitable evolution from vacuum energy to quark soup, nuclei and atoms, all the way to the emergence of life and self-awareness. But that does not explain where space, time and the laws came from, or why there is something rather than nothing.

I have quibbles with Baggott's book. He gives a dated picture of inflation (tying it to symmetry breaking), gets the temperature of the cosmic microwave background wrong (it is 2.7255 kelvin) and calls the lumpiness that led to the formation of cosmic structures anisotropy, rather than inhomogeneity. But these gaffes do not interfere with the larger narrative and are a by-product of his sweeping scope and detailed description.

The longing to understand our place in the cosmos is universal. Baggott and Randall lay out how much of the story we understand, and how interconnected it all is. They remind us that big questions remain in this most wonderful scientific adventure. ■

Michael S. Turner is professor of astronomy and astrophysics, and of physics at the University of Chicago, Illinois, and director of the Kavli Institute for Cosmological Physics. e-mail: mturner@kicp.uchicago.edu

NEUROSCIENCE

The mechanics of mind

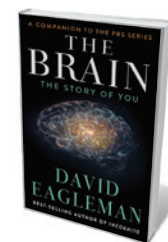
Daniel Bor enjoys a sophisticated study of how the meat in our skulls generates the self.

In my bolder moments, I consider neuroscience to be one of the most fundamental scientific fields. The brain is, after all, the location of our experiences and identities, and our main tool for understanding every facet of the Universe. *The Brain* by neuroscientist David Eagleman ambitiously promotes this view. Built around a series of fundamental questions, such as "what is reality?", it calls on a wide range of classic and recent findings, including innovative experiments by Eagleman himself, to demonstrate how brain science is optimally placed to answer those questions.

Eagleman begins by arguing that the

brain determines who we are, and how we change. He illustrates just how dramatic such changes can be through the case of Charles Whitman, who in the 1960s switched from mild-mannered bank clerk to violent murderer because of a small tumour pressing on his amygdala, an area of the brain linked to aggression and fear.

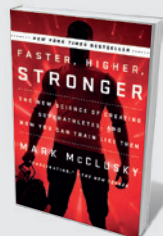
Although the brain's development has a disproportionate role in human identity, with synaptic pruning in infancy a key shaping factor, our brains remain plastic throughout our lives. Eagleman demonstrates this with the well-known example of London taxi drivers found to have enlarged



The Brain: The Story of You
DAVID EAGLEMAN
Pantheon: 2015.

hippocampi — key to memory consolidation — after memorizing thousands of the capital's streets. Memory is the bedrock of our identities, but Eagleman highlights how the past is very much a reconstruction bordering on mythology. A case in point is

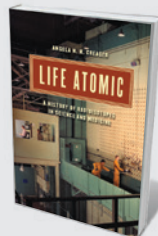
the relative ease with which false memories can be implanted. The emerging picture is far removed from one ►



Faster, Higher, Stronger

Mark McClusky (Plume, 2015)

From the primitive "bag-and-valve" apparatus used to measure runner's oxygen intake in the 1920s to today's Silicon Valley performance labs, Mark McClusky shows how sports science has helped humans to push their physical limits, and why we keep striving to beat the best.



Life Atomic: A History of Radioisotopes in Science and Medicine

Angela N. H. Creager (Univ. Chicago Press, 2015)

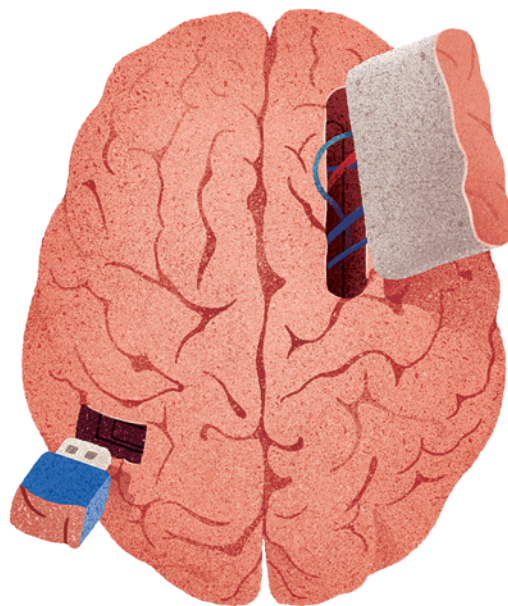
Radioisotope by-products of atomic energy are vital to molecular biology. Historian Angela Creager archives atoms, from carbon-14 and its role in studying photosynthesis to slow-decaying iron-59, which traces nutrients metabolizing in the body.

► in which we have a single personal identity. Instead, he notes, “from cradle to grave, we are works in progress”.

Everyone’s world view, then, is fallible. Eagleman extends that idea by focusing on perception. We tend to assume that we have a direct connection to what is out there, and that all that we experience is all there is, but the truth is very different. He writes how “every sight, sound, smell — rather than being a direct experience, is an electrochemical rendition in a dark theater”. But we are not experiencing a continuous flow of sight and sensation. We sample the world through saccades — jumping visual snapshots. From these, the brain constructs a continuous narrative heavily biased by expectations. We believe this narrative implicitly, even when it goes horribly wrong, such as in schizophrenic delusions.

In light of the idea that our view of our own minds is also deeply skewed, Eagleman challenges the primacy of consciousness. With most forms of expertise, such as driving, the conscious mind merely gets in the way. Furthermore, unconscious influences on our decisions are pervasive. For instance, judges deciding whether to award parole to prison inmates are more likely to do so if they have eaten beforehand. Decisions are little more than the product of unconscious neural battles between competing drives. Eagleman uses this stance to argue for an end to the catastrophic US war on drugs that began in the 1970s, and to call for more sympathy for, and (neuroscientific) understanding of, the plight of people addicted to drugs. He is helping addicts to reduce cravings by getting them to retrain their brain activity through neurofeedback in a functional magnetic resonance imaging scanner. They can view real-time summaries of the relative activities of their “craving” and “suppression” brain networks, and can practise strategies to discover the most effective way of suppressing cravings.

What of our relationships with others? Here, Eagleman notes that “what we demarcate as you is simply a network in a larger network”. He describes how social exclusion can, like physical harm, activate pain centres in the brain such as the insula. Empathy,



neurally speaking, invokes emotions as if we were experiencing for ourselves the events that we see others experience. However, when we consider members of “outgroups” to which we have no social ties, our empathic and social neural responses are flattened — as if we were dehumanizing them. This, Eagleman argues, is the neural mechanism that allows us to switch from being friendly to neighbours to wanting to wipe out their entire ethnic group, as for instance happened in the Bosnian war of the 1990s. Eagleman suggests that education about the neural underpinnings of our responses to outsiders is key to reducing the chance of genocides.

Eagleman ends by considering the future of humanity, and how neuroscience can technologically reshape almost every aspect of our lives. Although by far the most speculative part of the book, this is also the most fascinating. Eagleman describes the senses as flexible “peripheral plug-and-play” devices, with the brain not caring what input it receives as long as it is useful. We already exploit this feature with cochlear and retinal implants for people who have hearing or visual impairments. How much further can we take it?

Eagleman and his graduate student Scott Novich have developed an electronic vest that provides tactile feedback to the torso

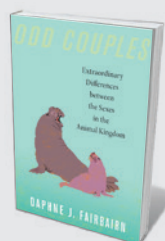
through arrays of small vibrating motors, and are testing it on people with impaired hearing, to allow them to ‘hear’ through touch.

The vest could be used for almost any real-time information stream, such as weather, stocks or altitude readouts in plane cockpits. And enhancing senses is only half the story: if we can control a robot arm through motor-cortex activity (see *Nature* **497**, 176–178; 2013), could someone check their e-mails while their brain-computer interface manages a vacuum cleaner?

Throughout, Eagleman provides multiple, varied explanations for what consciousness is and what it is for; he settles on neuroscientist Giulio Tononi’s integrated information theory. This equates high levels of consciousness with information that is widespread throughout a network capable of supporting many different information states. Tononi’s theory is consistent with the possibility of uploading our minds into computers. Without being limited by our fragile biology, we might feasibly travel to extrasolar worlds, “pausing” the computer simulation of our minds on the bulk of the journey to avoid boredom. Although such ideas are immensely fun to imagine, to computationally capture our brains we would have to be able to read every cellular detail of this incredibly complex organ — a feat that is centuries away, if it will ever be possible.

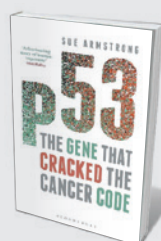
With such exciting themes, *The Brain* — a companion book to Eagleman’s upcoming six-part television series of the same name on the US Public Broadcasting Service — is an ideal introduction to how biology generates the mind. Readers familiar with this field will be revisiting a range of classic research, and might feel frustrated that more depth is not given in places. The science, however, is structured around crucial and wide-ranging questions, saturated with personal and social relevance. And Eagleman’s answers are consistently clear, engaging and thought-provoking. ■

Daniel Bor is a neuroscientist at the University of Sussex in Brighton, UK, and is the author of *The Ravenous Brain* and co-author of *30-Second Brain*.
e-mail: danielxbor@gmail.com



Odd Couples

Daphne J. Fairbairn (Princeton Univ. Press, 2015)
As biologist Daphne Fairbairn observes, males and females of one species can differ greatly in colour, size and shape. Blanket-octopus females, for instance, outgrow males by 2 metres, quashing the dominant-male stereotype (see Suzanne Alonzo’s review: *Nature* **496**, 427–428; 2013).



p53: The Gene that Cracked the Cancer Code

Sue Armstrong (Bloomsbury Sigma, 2015)
From its discovery in 1979 to its current place in cutting-edge gene therapy, p53 is the most studied gene in history. As Sue Armstrong details in this chronicle of genetics derring-do, its crucial role is to protect us from cancer, and the future of tumour treatment could depend on it. *Emily Banham*

Correspondence

US 'openness' bill is counterproductive

The US 2015 Secret Science Reform Act puts transparency in environmental science under the spotlight (D. Sarewitz *Nature* 525, 159; 2015). But the bill could end up weakening, rather than strengthening, environmental laws.

Of course, whatever scientific data can be made public should be. However, the bill as written will prevent agencies from using the best available science to protect public health. It stipulates that the US Environmental Protection Agency (EPA) should adopt new rules only after making raw data on pollution — including its effects on individuals — publicly available. The catch-22 is that the agency is rightly prohibited from revealing any such information that is confidential.

The EPA and scientific organizations have repeatedly raised these concerns, but the bill's sponsors have not addressed them. I agree with critics who conclude that its real intention could be to weaken environmental laws. Those politicians who hide their policy preferences behind scientific arguments must be held accountable.

Andrew A. Rosenberg *Union of Concerned Scientists, Cambridge, Massachusetts, USA.*
arosenberg@ucsusa.org

Community initiative tackles urban heat

A community-driven initiative is amassing data on urban form and function to help cities worldwide to develop their own heat-adaptation plans (see *Nature* 524, 402–404; 2015). It is called the World Urban Database and Access Portal Tools (WUDAPT; www.wudapt.org).

Urban experts use freely available Landsat satellite imagery to create and study local climate zones across their city

(see also B. Wake *Nature Clim. Change* 2, 487; 2012). Citizen scientists gather information on building materials and function, landscape morphology and vegetation types.

These extensive open-access WUDAPT data will provide a standardized characterization of the world's major cities and detailed input data for urban climate models. City planners and climate modellers will be equipped with accurate boundary conditions for investigating a range of mitigation and adaptation scenarios. Other applications include modelling the effects of changes to the energy infrastructure and improving the assessment of greenhouse-gas emissions through better calculations.

Linda See *International Institute for Applied Systems Analysis, Laxenburg, Austria.*

Gerald Mills *University College Dublin, Ireland.*

Jason Ching *University of North Carolina, Chapel Hill, USA.*
see@iiasa.ac.at

New oil investments boost carbon lock-in

Bringing new oil supplies to market could have an unexpectedly large impact on global emissions (see P. Erickson and M. Lazarus *Nature Clim. Change* 4, 778–781; 2014). New sources of oil increase carbon dioxide emissions in the short term, and make it harder and more expensive to scale down production in the long term. This 'carbon lock-in' entrenches our dependence on fossil fuels and commits economies to higher emissions (see go.nature.com/djtala).

The scale of investment in oil supplies and the profits they bring over the long term already dwarf those associated with other fuels. And significant barriers still confront the adoption of competing low-carbon technologies, such as electric vehicles. Capital-intensive oil

fields that have low operating costs relative to oil prices — as with most offshore oil deposits — make it even harder for us to switch. To wean us off oil, world leaders urgently need to curtail the billions of dollars that are currently earmarked for oil exploration and extraction.

Peter Erickson, Michael Lazarus *Stockholm Environment Institute, Seattle, Washington, USA.*

pete.erickson@sei-us.org

Ban unfair pricing of equipment imports

We believe that imported scientific equipment should be more fairly priced for researchers in China. We can be charged up to three times more than our counterparts in Western laboratories. By contrast, price differences for imported equipment are insignificant for scientists in Japan and Australia.

Examples of overpriced equipment include lasers for measuring distance, Doppler ultrasonic flow meters, data loggers and soil- or water-related sensors. Many of China's researchers are currently forced to spend as much as 60% of their funding on foreign equipment.

Our informal consultations with researchers and sales representatives suggest that prices in the Chinese market are inflated by distributors' charges and service fees, deals between distributors and manufacturers, and for other unknown reasons. Shipping fees account for only a small proportion of the surcharge.

To further its research and development, China needs to introduce more powerful regulations against these pricing monopolies. It should also step up its own manufacturing capability of key scientific equipment.

Rengui Jiang *Xi'an University of Technology, Xi'an, China.*
jrengui@163.com

Continental-drift opus turns 100

One hundred years ago this year, the legendary German explorer, geophysicist and meteorologist Alfred Wegener published his milestone book *The Origin of Continents and Oceans* (see *Nature* 127, 861; 1931). His theory of continental drift was initially viewed as heresy by the scientific community, yet his book was later translated into many languages and updated regularly until 1929.

For his opus, Wegener assembled an array of geological, palaeontological and geophysical data. They are best explained, he argued, by hypothesizing that major landmasses eventually broke apart and went their separate ways. After his death, his ideas were largely forgotten until the 1960s, when geophysicists demonstrated the phenomenon of sea-floor spreading (see N. Oreskes *Nature* 501, 27–29; 2013). Plate tectonics has since gained acceptance as a synthetic theory with huge explanatory power.

Wegener died in 1930 while exploring in Greenland. Buried in the ice, his body has sailed westwards at a rate of about 2 centimetres per year on the back of the North American plate. He would have been glad to know that it will have travelled some 20 kilometres in a million years' time — in accordance with his visionary theory.

Marco Romano *Sapienza University of Rome, Italy.*

Richard L. Cifelli *Sam Noble Oklahoma Museum of Natural History, Norman, Oklahoma, USA.*

rich.cifelli@gmail.com

CONTRIBUTIONS

Correspondence may be submitted to correspondence@nature.com after consulting the author guidelines at <http://go.nature.com/cmchno>.

Testing the mid-latitude hydrologic seesaw

ARISING FROM: K.-N. Jo *et al.* *Nature* **508**, 378–382 (2014); doi:10.1038/nature13076

The hydrologic cycle is one of Earth's fundamental physical processes. Jo *et al.*¹ propose that a centennial to millennial scale interhemispheric hydrologic seesaw operates in the mid-latitudes; this is an important concept because, if correct, it provides further insight into the operation of the cycle. The hypothesis is tested here, but results lead to questions concerning the general validity of the proposed seesaw. Thus further testing is required to prove the concept. There is a Reply to this Brief Communication Arising by Jo, K.-N. *et al.* *Nature* **526**, <http://dx.doi.org/10.1038/nature14977> (2015).

In high and low latitudes centennial to millennial scale interhemispheric climatic seesaws have been identified with respect to both temperature^{2,3} and precipitation⁴, but in the mid-latitudes there is uncertainty regarding the expression and strength of any interhemispheric seesaw effect, although flooding episodes may be antiphase⁵.

In the mid-latitude monsoonal environment of Korea, Jo *et al.*¹ showed the probability density distribution of speleothem ages to be a reflection of changes in climate: high and low growth frequency phases were found to be well-matched to interglacial/interstadial and glacial/stadial periods, respectively, with faster growth associated with warmer intervals linked to the summer monsoon, the strength of which is sensitive to the position of the Intertropical Convergence Zone (ITCZ). Jo *et al.*¹ considered their data to support the idea that shifts in the ITCZ cause latitudinal displacements of all climatic zones, although another consideration is that changes in the latitudinal temperature gradient cause expansion and contraction of the westerly wind circulation. As a means of delineating the interhemispheric impacts of ITCZ migration, Jo *et al.*¹ compared the growth record of Korean speleothems with a set of speleothems from southeastern Australia⁶. This showed the records to be inversely correlated, and this was considered by Jo *et al.*¹ to provide evidence for an interhemispheric hydrologic seesaw. But since the Australian records are also inversely correlated with those from nearby New Zealand (Fig. 1), the situation must be more complex. The representativeness of the data sets for each hemisphere is also an issue.

If this interhemispheric hydrologic process were in general operation, then the Korean record should also be inversely correlated with other speleothem series from the southern mid-latitudes. This is tested here by comparing the Korean record with a probability density

distribution of speleothem ages from New Zealand⁷. The records are imperfectly matched, but despite that the timing (within a few millennia) of most speleothem growth peaks tends to be broadly in-phase rather than inversely correlated (Fig. 1), both records displaying enhanced growth in warm interstadials, but reduced growth in cool stadials, conditions during these intervals in New Zealand being described in ref. 8. Several growth peaks (30–40, 49–56, 70–74, 75–85, 98–105, 115–125 ka) identified in the mid-latitude European speleothem record^{9,10} also broadly coincide with peaks in New Zealand. Consequently, since speleothem growth in New Zealand is out-of-phase with southeastern Australia, yet shows a tendency to be in-phase with Korea, a simple mid-latitude interhemispheric hydrologic seesaw is not supported. The situation appears more complex.

So what explanation can be offered for the Australian record^{6,11} being out-of-phase with New Zealand, Korea and Europe? The main control of speleothem growth in semi-arid southeastern Australia is water supply⁶. Like a tap, it turns off speleothem growth when evapotranspiration exceeds precipitation. This occurs under interstadial and interglacial conditions when the mean position of rain-bearing westerlies moves poleward. This factor is less critical in oceanic New Zealand, Korea and western Europe, because these regions usually have an annual water surplus, with deficits only sometimes occurring under glacial and stadial conditions.

In conclusion, the hypothesis of a centennial to millennial scale mid-latitude interhemispheric hydrologic seesaw is not supported by available speleothem evidence. Hence doubts are raised concerning its general validity. The pattern of speleothem growth in Korea is related to monsoonal rains, but this factor is less directly applicable in the other areas considered, where more important is the effect of expansion and contraction of the westerly circulation. Meridional shifts of the mid-latitude tropospheric jet stream associated with the southern annular mode is likely to be the main driver of speleothem growth patterns in the southern mid-latitudes, because it has a major influence on regional climates^{12–14} with contrasting effects in southern Australia (on-off tap) and New Zealand (temperature and humidity driven growth frequency and rate changes). Further testing of the seesaw hypothesis is needed, ideally comparing monsoon regions in both hemispheres.

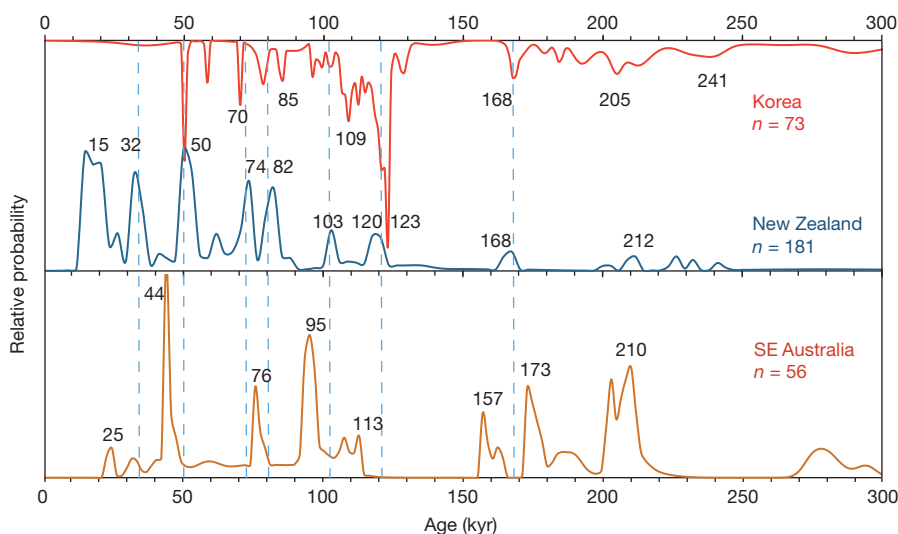


Figure 1 | Comparison of the relative probability density distributions of speleothem ages from Korea, southeastern Australia and New Zealand for the interval 11–300 ka. Numbers show the ages of peaks in the distributions. Dashed vertical lines indicate where growth peaks in Korean or New Zealand speleothems coincide with troughs in the Australian distribution. Data sets and analytical methods used are explained in Methods.

Methods

The New Zealand speleothem data set used here is a modification of that presented in ref. 7. First, data were limited to between 11 ka and 300 ka in order to match the age range addressed by Jo *et al.*¹ and, second, new speleothem ages were added from ref. 15. This accounts for differences observed in the probability density function curves for 0–600 ka (ref. 7) and 11–300 ka (Fig. 1). The Australian data set used by Jo *et al.*¹ was supplemented by additional age data for the same region from ref. 11. Probability density analysis followed routines developed in ref. 16.

Paul W. Williams¹

¹University of Auckland, Auckland 1142, New Zealand.
email: p.williams@auckland.ac.nz

Received 12 February; accepted 13 August 2015.

1. Jo, K.-N. *et al.* Mid-latitude interhemispheric hydrologic seesaw over the past 550,000 years. *Nature* **508**, 378–382 (2014).
2. Barker, S. *et al.* Interhemispheric Atlantic seesaw response during the last deglaciation. *Nature* **457**, 1097–1102 (2009).
3. EPICA Community Members. One-to-one coupling of glacial climate variability in Greenland and Antarctica. *Nature* **444**, 195–198 (2006).
4. Cheng, H., Sinhua, A., Wang, X., Cruz, F. W. & Edwards, R. L. The global paleomonsoon as seen through speleothem records from Asia and the Americas. *Clim. Dyn.* **39**, 1045–1062 (2012).
5. Macklin, M. G., Fuller, I. C., Jones, A. F. & Bebbington, M. New Zealand and UK Holocene flooding demonstrates interhemispheric climate asynchrony. *Geology* **40**, 775–778 (2012).
6. Ayliffe, L. K. *et al.* 500 ka precipitation record from southeastern Australia: evidence for interglacial relative aridity. *Geology* **26**, 147–150 (1998).

7. Williams, P. W., Neil, H. L. & Zhao, J.-X. Age frequency distribution and revised stable isotope curves for New Zealand speleothems: palaeoclimatic implications. *Int. J. Speleol.* **39**, 99–112 (2010).
8. Williams, P. W., McGlone, M., Neil, H. & Zhao, J.-X. A review of New Zealand palaeoclimate from the Last Interglacial to the global Last Glacial Maximum. *Quat. Sci. Rev.* **110**, 92–106 (2015).
9. Baker, A., Smart, P. L. & Ford, D. C. Northwest European palaeoclimate as indicated by growth frequency variations of secondary calcite deposits. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **100**, 291–301 (1993).
10. Hercman, H. Reconstruction of paleoclimatic changes in central Europe between 10 and 200 thousand years BP, based on analysis of growth frequency of speleothems. *Stud. Quat.* **17**, 35–70 (2000).
11. Moriarty, K. C., McCulloch, M. T., Wells, R. T. & McDowell, M. C. Mid-Pleistocene cave fills, megafaunal remains and climate change at Naracoorte, South Australia: towards a predictive model using U-Th dating of speleothems. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **159**, 113–143 (2000).
12. Kidston, J., Renwick, J. A. & McGregor, J. Hemispheric-scale seasonality of the southern annular mode and impacts on the climate of New Zealand. *J. Clim.* **22**, 4759–4770 (2009).
13. Hendon, H. H., Thompson, D. W. J. & Wheeler, M. C. Australian rainfall and surface temperature variations associated with the Southern Hemisphere annular mode. *J. Clim.* **20**, 2452–2467 (2007).
14. Meneghini, B., Simmonds, I. & Smith, I. N. Association between Australian rainfall and the Southern Annular Mode. *Int. J. Climatol.* **27**, 109–121 (2007).
15. Whittaker, T. E., Hendy, C. H. & Hellstrom, J. C. Abrupt millennial-scale changes in intensity of Southern Hemisphere westerly winds during marine isotope stages 2–4. *Geology* **39**, 455–458 (2011).
16. Ludwig, K. R. *User's Manual for Isoplot 3.6: A Geochronological Toolkit for Microsoft Excel*. Special Publication 4 (Berkeley Geochronology Center, 2008).

Competing Financial Interests Declared none.

doi:10.1038/nature14976

Woo *et al.* reply

REPLYING TO P. W. Williams *Nature* **526**, <http://dx.doi.org/10.1038/nature14976> (2015)

The hypothesis of a mid-latitude interhemispheric hydrologic seesaw, based on relative probability data comparing speleothem growth between the Korean peninsula and southeastern Australia¹, has been questioned by Williams in the accompanying Comment². This questioning was based on the assumption that all the palaeo-hydroclimatic data from all the mid-latitude regions could be equally applied. However, we consider that this approach is not appropriate because the seesaw pattern should not be shown as straight lines exactly parallel to latitude. It can show sinuous and sometimes isolated patterns, depending on geographic location and hydrologic regime at different timescales³. In addition, our hypothesis is also supported by other New Zealand data from elsewhere⁴.

Our hypothesis stated that an interhemispheric hydrologic seesaw could be interconnected globally up to mid-latitude regions throughout interglacial–glacial cycles (that is, on orbital timescales)¹. Williams² claims that our hypothesis should be challenged unless seesaw patterns in all the mid-latitude regions including New Zealand can be equally applied. However, this approach is not appropriate because even tropical to subtropical hydro-climatic regions directly affected by the migration of the ITCZ also show highly sinuous global geometry³. In fact, the seasonal position of the ITCZ expands to higher latitudes on the sides of the continents, and the overall geometry of the ITCZ is affected by the distribution of land-masses^{3,5}. This may well imply that the interhemispheric seesaw has reached more effectively during glacial time to continental southeastern Australia, but not to maritime New Zealand. The pluvial events from southeastern Australia support our hypothesis, and reflect the intensity changes in the Walker circulation or the Australian monsoon, which are directly related to the position of the ITCZ⁶, but not

closely to the position of New Zealand. Also, if the past positions of westerlies were responsible for Australian speleothem growth in the past, they should have been intimately interconnected with ITCZ movements^{7,8}.

We consider that the argument put forward by Williams² does not support the other stalagmite data in New Zealand⁴. Isotopic and growth rate data of a stalagmite from South Island, New Zealand, show clear antiphase patterns with East Asian monsoon records over millennial to orbital timescales. This strongly suggests that East Asian and New Zealand regions have opposite patterns in orbital scale rainfall intensity. Also—considering a centennial to millennial scale which was included in Williams' arguments but not in ours—flooding events reconstructed by fluvial sedimentary sequences in New Zealand show an anti-phase trend with UK data in the Northern Hemisphere⁹.

Based on the data provided by Williams², it appears that the growth frequency patterns of New Zealand speleothems are neither in phase nor out of phase with those of the Korean data. It should be noted that New Zealand speleothems grew very actively, even during the Last Glacial Maximum (LGM), and continued to grow throughout entire glacial and interglacial periods¹⁰. The LGM peak should not be overlooked if both data sets are claimed to be broadly in-phase on an orbital scale. It is clear that the Korean speleothem growth frequency decreases from the Eemian to the LGM, which is also supported by the number of speleothem growths¹, however this trend cannot be identified in New Zealand data. We think that this opposite trend may well be the possible orbital scale contrast. Even though some peaks share similar age intervals, exact timings of those peaks are difficult to resolve on a millennial timescale within given error ranges and such a probability method¹. Thus it is difficult to determine the simple

relationship of speleothem growth patterns between the Korean peninsula and New Zealand.

In conclusion, we think that our hypothesis is worth testing, but we do not think that the New Zealand data provided by Williams² are sufficient to test the validity of our hypothesis. We hope that in the future, further analysed data from elsewhere could clarify this discrepancy.

Kyoung-nam Jo¹, Kyung Sik Woo², Sangheon Yi¹, Dong Yoon Yang¹, Hyoun Soo Lim³, Yongjin Wang⁴, Hai Cheng^{5,6} & R. Lawrence Edwards⁶

¹Korea Institute of Geoscience and Mineral Resources, Daejeon 305-350, South Korea.

²Department of Geology, Kangwon National University, Gangwondo 200-701, South Korea.

email: wooks@kangwon.ac.kr

³Department of Geological Sciences, Pusan National University, Busan 609-735, South Korea.

⁴College of Geography Science, Nanjing Normal University, Nanjing 210097, China.

⁵Institute of Global Environmental Change, Xi'an Jiaotong University, Xi'an 710049, China.

⁶Department of Geology and Geophysics, University of Minnesota, Minneapolis, Minnesota 55455, USA.

1. Jo, K.-N. *et al.* Mid-latitude interhemispheric hydrologic seesaw over the past 550,000 years. *Nature* **508**, 378–382 (2014).
2. Williams, P. W. Testing the mid-latitude hydrologic seesaw. *Nature* **526**, <http://dx.doi.org/10.1038/nature14976> (2015).
3. Wang, X. *et al.* Interhemispheric anti-phasing of rainfall during the last glacial period. *Quat. Sci. Rev.* **25**, 3391–3403 (2006).
4. Whittaker, T. E., Hendy, C. H. & Hellstrom, J. C. Abrupt millennial-scale changes in intensity of Southern Hemisphere westerly winds during marine isotope stages 2–4. *Geology* **39**, 455–458 (2011).
5. Schneider, T., Bischoff, T. & Haug, G. H. Migrations and dynamics of the intertropical convergence zone. *Nature* **513**, 45–53 (2014).
6. Ayliffe, L. K. *et al.* 500 ka precipitation record from southeastern Australia: evidence for interglacial relative aridity. *Geology* **26**, 147–150 (1998).
7. Ceppi, P., Hwang, Y.-T., Liu, X., Frierson, D. M. W. & Hartmann, D. L. The relationship between the ITCZ and the Southern Hemispheric eddy-driven jet. *J. Geophys. Res. Atmos.* **118**, 5136–5146 (2013).
8. Toggweiler, J. R. Shifting westerlies. *Science* **323**, 1434–1435 (2009).
9. Macklin, M. G., Fuller, I. C., Jones, A. F. & Bebbington, M. New Zealand and UK Holocene flooding demonstrates interhemispheric climate asynchrony. *Geology* **40**, 775–778 (2012).
10. Williams, P. W., Neil, H. L. & Zhao, J.-X. Age frequency distribution and revised stable isotope curves for New Zealand speleothems: palaeoclimatic implications. *Int. J. Speleol.* **39**, 99–112 (2010).

doi:10.1038/nature14977

ASTROPHYSICS

Primordial stars brought to light

The earliest stars are of huge importance to the chemical history of the cosmos, but have previously existed only in theory. There is now strong evidence that such population III stars exist in the brightest galaxy yet found in the early Universe.

BETHAN JAMES

Stars of the first generation to form in the early Universe were essential to the growth and evolution of structure in the cosmos. But these primordial population III stars, which are thought to reside in the youngest galaxies, have been notoriously elusive. Writing in *The Astrophysical Journal*, Sobral *et al.*¹ present the first observational evidence for the existence of these stars, in the brightest galaxy so far detected from the end of the 'reionization' era (when the Universe transitioned from a neutral to an ionized state). Moreover, the authors infer that this select population exists alongside an evolved stellar group, suggesting that it may be easier to detect systems containing population III stars than had been thought.

During the first 200 million years after the Big Bang, only hydrogen, helium and very small amounts of lithium existed. The first stars were thus formed from chemically pristine gas clouds and were 'metal free' — that is, they did not contain elements heavier than hydrogen and helium. Because of their primordial composition, these stars grew to become 1,000 times more massive than the Sun (according to some models) and were exceptionally hot, emitting copious amounts of ionizing radiation^{2,3}. Consequently, theoretical models predict that population III stars should have distinct spectral signatures, such as strong hydrogen and helium emission lines (the Lyman- α line in the case of hydrogen), but no metal emission lines^{4–6}. These stars are of enormous cosmological importance: they are prime candidates for reionizing⁷ the intergalactic medium and producing the first metals that chemically enriched subsequent stellar generations and the Universe as a whole.

Despite their strong theoretical background, population III stars have remained undetected for several possible reasons. First, the high masses and short life spans⁵ typically expected of these stars imply that they would now exist as obscure remnants (such as neutron stars or black holes), although low-mass members of the population would still be observable. Second, metals produced in the cores of population III stars may have been dredged



M. KORNMESSER/ESO

Figure 1 | Population III stars in the CR7 galaxy. Sobral *et al.*¹ present spectroscopic evidence for the existence of population III stars in CR7 (shown here as an artist's representation) — the most luminous galaxy known in the early Universe. These primordial stars were born from clouds of gas that contained hydrogen and helium (and trace amounts of lithium), but no heavier elements.

up to their surfaces, causing them to seem less metal-poor than they initially were. Finally, first-generation stars may exist in only the earliest, and hence most-distant galaxies, which are beyond the reach of current telescopes (although it has been predicted⁸ that galaxies containing population III stars might also have formed at later times in low-density regions of the Universe).

Detecting distant (high-redshift) galaxies is extremely difficult, because the sources are faint and only their ultraviolet (UV) radiation is observable, after being redshifted to longer optical and infrared wavelengths. Furthermore, Lyman α , the strongest UV spectral line emitted from these high-redshift systems, is easily attenuated through absorption by interstellar dust and neutral hydrogen atoms. Astronomers carry out deep-imaging surveys to look for these galaxies (known as luminous

Lyman- α emitters), because such systems are excellent candidates for harbouring population III stars and can provide insights into the epoch of reionization.

In one such survey, Sobral and his team previously found⁹ a particularly luminous candidate source called CR7 (Fig. 1). In their present follow-up optical and infrared spectroscopic observations, Sobral *et al.* report that not only is this the most intrinsically luminous Lyman- α emitter found at the reionization era, but also that this system shows promising population III spectral signatures. In addition to exceptionally strong Lyman- α emission, the authors detect an emission line of ionized helium, but no other features — this suggests a metal-free galaxy.

The authors find that the observations of CR7 are best explained by a hybrid stellar population containing both young, metal-free

stars and an older, chemically evolved population, respectively emitting most of the UV and optical light. Sobral *et al.* corroborated these model-based predictions using images obtained by the Hubble Space Telescope which show that CR7 consists of three spatially separated clumps.

This structure is consistent with a theoretical study¹⁰ that proposed the 'inside-out' formation of population III stars. The authors' observations suggest that a wave of star formation swept over CR7, progressively moving outwards from the old red clumps towards the young, UV-bright clump 5 kiloparsecs away, where population III stars are inferred to exist. This distance is large enough to prevent contamination of the UV-bright clump by metals created in the structure's interior. According to this picture, photons from the old stars not only helped to ionize a large bubble around them, but also prevented any surrounding gas from forming stars for some time by impeding the gas's gravitational collapse. This created an ideal environment for population III stars to eventually form in the chemically pristine clump, while allowing the Lyman- α photons to escape because there was no neutral hydrogen to absorb them. This formation mechanism implies that, in Lyman- α emitters, population III stars would most probably be detected in hybrid stellar populations.

Sobral and colleagues' discovery of this population III system is not without caveats. For example, to match the ratio of helium-to-Lyman- α line emission predicted by models, the authors deduce that 75% of the Lyman- α emission must have been lost through scattering or absorption. In addition, the proposed scenario in which evolved, 'second-generation' stars can ionize a region without chemically polluting it, while also conveniently holding back star formation in the vicinity until population III stars arise, seems somewhat ad hoc, although it is not impossible. Finally, the authors state that the properties of CR7 can also be explained by gas falling into a black hole formed from the direct collapse of primordial gas, another theoretical construct. However, material around black holes produces broad emission-line profiles and X-ray radiation, neither of which are detected in the current observations. Deep observations with next-generation X-ray telescopes might be able to distinguish between the population III and black-hole scenarios¹¹.

Overall, Sobral *et al.* present the most promising observational evidence for the existence of population III stars found so far. An immediate implication of the CR7 finding is that these stars may be more readily detectable than previously thought. Rather than existing solely in isolation in the earliest galaxies, metal-free stars may also form alongside evolved stellar populations. Therefore, all luminous Lyman- α emitters are excellent candidates for harbouring population III stars and, as such, are ideal

targets for the James Webb Space Telescope, the next-generation infrared observatory. This telescope's unprecedented sensitivity will enable it to detect both bright and faint metal emission lines in the optical spectra of distant galaxies, and to confirm whether CR7 and similar systems contain the elusive population III stars. ■

Bethan James is at the Institute of Astronomy, University of Cambridge, Cambridge CB3 0HA, UK.
e-mail: bjames@ast.cam.ac.uk

NANOTECHNOLOGY

Platelet mimicry

Cloaking drug-loaded nanoparticles with platelet membranes enhances the drugs' abilities to target desired cells and tissues. This technology might improve treatments for cardiovascular and infectious diseases. [SEE LETTER P118](#)

OMID C. FAROKHZAD

The development of nanoparticles that can carry drugs to target sites in the body promises safer and more-effective drug delivery to solve myriad medical problems. It has proved difficult to create the complex exterior surface that allows these nanocarriers to undergo 'normal' biological interactions^{1,2}, but, by turning to nature for design cues, scientists have begun to develop such biomimetic nanoparticles^{3,4}. On page 118 of this issue, Hu *et al.*⁵ report that nanoparticles coated with the membrane of blood platelets are shielded from the body's immune responses, and possess platelet-like binding properties that allow them to

target desired cells and tissues. Alongside broad therapeutic implications, the study blurs the line between materials science and biochemistry, introducing techniques that could benefit both nanoengineering and biomembrane research.

When blood vessels are damaged, the injury exposes proteins such as collagen that are abundant in the subendothelial layer underneath the vessel lining. Platelets — small, non-nucleated membrane-bound cell fragments circulating in the blood — bind to these proteins with strong affinity and then release blood-clotting factors, promoting the formation of a platelet plug that helps to heal the wound. Because many conditions, including cancer, inflammation and trauma, are associated with vascular

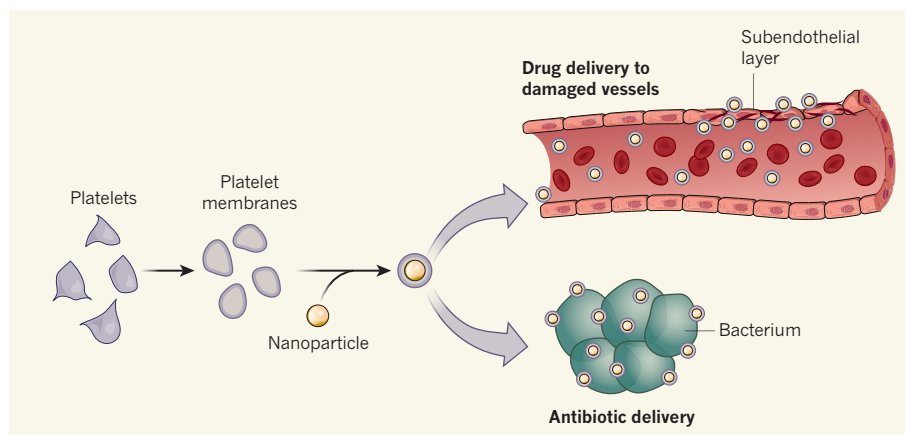


Figure 1 | Designing nanocarriers. Hu *et al.*⁵ have developed nanocarriers that improve drug delivery to desired targets. The authors isolated the membranes of platelets from human blood and used them to cloak synthetic nanoparticles loaded with drug. These nanocarriers mimic the biological properties of platelets, enabling them to evade immune detection in the body. They can bind to the exposed subendothelial layer of damaged vessels, improving drug delivery for many diseases associated with damaged vasculature, and can also improve antibiotic delivery to bacteria in the body. (Adapted from ref. 5.)

damage, platelets have long inspired drug-delivery research. Nanoparticles have been engineered to display platelet-like ligands on their surface, which facilitates binding to subendothelial components^{6,7}. In addition, platelet morphology and clotting mechanisms have been modelled, with the aim of enhancing drug targeting^{7,8}. However, such efforts have failed to produce nanoparticles that can truly mimic the behaviour of platelets.

Platelets have also attracted interest in studies of infectious disease, because several bacterial species express surface proteins that interact with platelet receptors. This platelet–bacterium interaction has been linked to lethal complications during infection⁹. For instance, the high volume and velocity of blood that passes through the heart valves make that area susceptible to injury, and, in infective endocarditis, invasive microbes adhere to injured valve surfaces and promote further platelet aggregation. This is a serious therapeutic challenge, because the clot-encased microbes at the valve are inaccessible to antibiotic treatment and evade the immune response. Without effective intervention, approximately 40% of hospital patients who contract infective endocarditis will die¹⁰.

The researchers behind the current study have previously developed nanoparticles coated with the membranes of red blood cells and cancer cells, and have shown that these nanoparticles can be used to neutralize bacterial toxins and for anti-cancer vaccinations, respectively^{11–13}. Building on this success, Hu *et al.* developed polymeric nanoparticles coated in platelet membrane that mimic many of the biological functions of platelets (Fig. 1). Imbuing the nanoparticles with platelet-like properties was a notable challenge, but the authors took advantage of the fact that there is a differential charge distribution between the outer and inner surface of the platelet membrane, due to the abundance of negatively charged sialic acid molecules on the outer surface. Hu and colleagues made their nanoparticles negatively charged and so, through electrostatic-charge repulsion with the platelets' outer membranes, the nanoparticles preferentially bound to the inner membrane. This ensured that the membrane was 'right-side-out' on the nanoparticle surface.

Hu and co-workers' nanocarriers have a more complete set of membrane proteins than previous platelet-mimicking nanoformulations — the nanoparticles were coated in 15 immunomodulatory and subendothelial-binding components. This membrane cloak enabled the particles to bind effectively to human collagen in *in vitro* assays, and to target regions of damage in isolated blood vessels. The authors demonstrated that the nanocarriers successfully evaded detection by immune cells, and were well tolerated by rodents.

The narrowing of arteries or valves as a result of excessive cell proliferation can pose

problems following corrective surgery. Hu and colleagues' nanoformulation effectively prevented vessel thickening in a rat model of this disorder, which is known as restenosis. The nanoparticles selectively bound to injured arteries, enabling the sustained release of an antiproliferative drug.

Perhaps more exciting is the nanoparticles' ability to target bacterial species that adhere to platelets. Targeted antibiotic delivery is a major research topic given the rising threat of antibiotic resistance. However, identifying an injectable and broadly applicable pathogen-targeting particle has been a technical hurdle to developing antibacterial nanocarriers. The authors showed that their technology could overcome this challenge for the bacterium *Staphylococcus aureus*, a common pathogen. Compared with free antibiotic, the nanoparticles improved delivery of antibiotics to the bacteria both *in vitro* and in infected mice. This ability to specifically target bacteria might enable platelet-membrane-coated nanoparticles to tackle severe complications of infection, such as the presence of bacteria in the blood, which can cause sepsis and the spread of infection. And by directing higher drug doses to the pathogen, these nanoparticles offer the hope of boosting the effectiveness of antibiotics whose efficacy is on the wane.

Given the innovative nature of Hu and colleagues' nanocarriers, manufacturing and regulatory standards must be established before they can be used in the clinic. The past decade has seen considerable advances¹ in establishing best practice in this area — there have been improvements in the processing of human blood products to enhance their

preservation and function, and complex synthetic nanocarriers have been engineered and used in human clinical trials. Extra risks must be taken into account when designing nanocarriers that combine biological and synthetic components, but the biotechnology industry has the operating procedures in place to meet the required standards. These are exciting times in nanomedicine. The authors' biomimetic nanoparticles mark a new frontier, providing a glimpse into the future of the field. ■

Omid C. Farokhzad is in the Laboratory of Nanomedicine and Biomaterials, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. e-mail: ofarokhzad@bwh.harvard.edu

1. Kamaly, N., Xiao, Z. Y., Valencia, P. M., Radovic-Moreno, A. F. & Farokhzad, O. C. *Chem. Soc. Rev.* **41**, 2971–3010 (2012).
2. Nel, A. E. *et al. Nature Mater.* **8**, 543–557 (2009).
3. Cho, W. K. *et al. Proc. Natl Acad. Sci. USA* **109**, 21289–21294 (2012).
4. Pridgen, E. M. *et al. Sci. Transl. Med.* **5**, 213ra167 (2013).
5. Hu, C.-M. J. *et al. Nature* **526**, 118–121 (2015).
6. Kamaly, N. *et al. Proc. Natl Acad. Sci. USA* **110**, 6506–6511 (2013).
7. Anselmo, A. C. *et al. ACS Nano* **8**, 11243–11253 (2014).
8. Simberg, D. *et al. Proc. Natl Acad. Sci. USA* **104**, 932–936 (2007).
9. Fitzgerald, J. R., Foster, T. J. & Cox, D. *Nature Rev. Microbiol.* **4**, 445–457 (2006).
10. Prendergast, B. *Circulation* **121**, 1141–1152 (2010).
11. Hu, C. M., Fang, R. H., Copp, J., Luk, B. T. & Zhang, L. *Nature Nanotechnol.* **8**, 336–340 (2013).
12. Fang, R. H. *et al. Nano Lett.* **14**, 2181–2188 (2014).
13. Hu, C. M., Fang, R. H., Luk, B. T. & Zhang, L. *Nature Nanotechnol.* **8**, 933–938 (2013).

This article was published online on 16 September 2015.

PHENOLOGY

Spring greening in a warming world

Warmer temperatures have been associated with an earlier emergence of spring leaves each year. New data, however, suggest that leaf emergence is becoming less sensitive to temperature as global temperatures rise. SEE LETTER P.104

TREVOR F. KEENAN

For centuries, people have been fascinated by the timing of the arrival of spring, a season named for the 'springing forth' of the leaves of deciduous trees. It has long been known that spring leaf emergence is strongly linked to temperature^{1,2} — even in ancient Rome, Pliny the Elder realized that leaf emergence was a much better indicator of weather than were the constellations³. Leaves have emerged earlier over the past century,

as spring has become warmer. With global anthropogenic emissions currently exceeding previous worst-case scenarios⁴, considerable warming is expected in the coming decades. Will future warming lead to even earlier and greener springs? In this issue, Fu *et al.*⁵ (page 104) report results suggesting that the relationship between the seasonal timing of leaf emergence — spring phenology — and temperature is changing.

The relationship between spring temperatures and leaf emergence has allowed scientists



Figure 1 | Green and early. Temperature is the dominant factor in inducing the onset of spring leaf emergence in temperate deciduous forests. But Fu *et al.*⁵ suggest that factors such as reduced winter chilling are decreasing the sensitivity of spring leaf emergence to temperature.

of the Intergovernmental Panel on Climate Change to use changes in the timing of emergence as a key indicator of the ecological impact of climate change⁶. Apart from resulting in a greener spring, earlier leaf emergence affects various aspects of ecosystem function, and generates multiple feedbacks to the climate system⁷. It has thus been built into state-of-the-art Earth-system models, which predict a large advance in the timing of leaf emergence under future climate warming. To test the relationship between leaf emergence and warming, Fu and colleagues examined 33 years of observations of 7 forest species across 1,245 sites in Europe. Surprisingly, they discovered that spring leaf emergence has been getting less sensitive to temperature over time (Fig. 1). Their observation-based results call into question current model projections, and suggest that spring leaves might not emerge as early under future warming as had been previously expected.

Although it is generally accepted that temperature is the dominant driver of spring phenology in temperate deciduous forests, there is considerable uncertainty about the pathways of temperature's influence, with little agreement between models, experiments and observations^{8,9}. The timing of warming matters¹⁰, and the response seems to vary by species and perhaps by location or population¹¹. Many other factors could also play a part — primarily day-length (photoperiod) and winter-dormancy requirements, but also humidity and temperature variance.

Photoperiod has been shown to have a strong influence on some species, particularly on the *Fagus* (beech) genus, for which the effect of warm temperatures is limited if the days are too short¹².

Many species have also been shown to require a certain amount of chilling in winter before their release from dormancy¹³. This evolutionary mechanism, which is designed to prevent the costly damage a late frost can inflict on young leaves, ensures that winter has truly passed before leaves emerge. Changes in any of these factors could potentially modify the temperature response of leaf emergence, and explain the observed decline in temperature sensitivity reported by Fu and colleagues.

The authors tested three hypotheses to examine the potential underlying causes of their observations. They assessed the role of photoperiod, but could neither confirm nor rule out its influence. They also found no significant changes in the timing of leaf emergence due to temperature-variance changes, which suggests a limited role for this factor. The third hypothesis tested was that warmer winters had resulted in reductions in winter chilling, which could dampen the response of spring leaf emergence to a warmer spring. Using multiple models, the authors showed that declines in winter chilling could indeed lead to a lower temperature sensitivity, although the change in temperature sensitivity predicted by the models was considerably smaller than that observed.

This long-term trend of a decline in

winter chilling, in concert with a decline in the temperature sensitivity of spring leaf emergence, raises questions about the extent to which factors such as chilling requirements are already limiting the response of spring phenology to climate warming. However, the association falls short of a causal attribution, as the authors note, because temperature sensitivity was not observed to be markedly different in years that had more chilling than in years with less chilling. Furthermore, not all deciduous plants have chilling requirements, and many have low requirements that are met even under experimental warming¹⁴. For most species, the effect of chilling requirements is poorly understood.

The declining temperature sensitivity reported by Fu and colleagues is intriguing, but its root cause is still uncertain. More research is needed to assess whether other species and locations demonstrate a similar decline in temperature sensitivity — in particular, to examine the many other long-term records around the world, in combination with satellite observations of vegetation, experimental data and theoretical understanding.

Leaves emerge in spring as a result of responses that are hard-wired into the genetic code of trees. This might suggest that the response of phenology to environmental drivers should be highly predictable¹⁵, yet we are far from having a predictive science of phenology. Observations such as those presented by Fu *et al.*, which challenge models and contemporary understanding, go a long way towards getting us there. ■

Trevor F. Keenan is in the Department of Biological Sciences, Macquarie University, Sydney, New South Wales 2109, Australia. e-mail: trevor.keenan@mq.edu.au

1. De Réaumur, R. A. F. *Mém. Acad. R. Sci. Paris* 545–576 (1735).
2. Lieth, H. *Phenology and Seasonality Modeling* (Springer, 1974).
3. Bostock, J. & Riley, H. T. *The Natural History of Pliny* (Bohn, 1857).
4. Le Quéré, C. *et al. Nature Geosci.* **2**, 831–836 (2009).
5. Fu, Y. H. *et al. Nature* **526**, 104–107 (2015).
6. Pachauri, R. K. *et al. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (2014).
7. Richardson, A. D. *et al. Agric. Forest Meteorol.* **169**, 156–173 (2013).
8. Wolkovich, E. M. *et al. Nature* **485**, 494–497 (2012).
9. Richardson, A. D. *et al. Global Change Biol.* **18**, 566–584 (2012).
10. Friedl, M. A. *et al. Environ. Res. Lett.* **9**, 054006 (2014).
11. Parmesan, C. *Global Change Biol.* **13**, 1860–1872 (2007).
12. Zohner, C. M. & Renner, S. S. *New Phytol.* <http://dx.doi.org/10.1111/nph.13510> (2015).
13. Laube, J. *et al. Global Change Biol.* **20**, 170–182 (2014).
14. Fu, Y. H., Campioli, M., Deckmyn, G. & Janssens, I. A. *PLoS ONE* **7**, e47324 (2012).
15. Luo, Y., Keenan, T. F. & Smith, M. *Global Change Biol.* **21**, 1737–1751 (2015).

This article was published online on 23 September 2015.



50 Years Ago

My View of the World. By Erwin Schrödinger — This book consists of two long essays, hitherto unpublished ... In these pages Schrödinger puts forward a philosophical and indeed metaphysical view of the world, of human experience and of human nature ... Instead of thinking of different selves experiencing a common world, he invites us to think of a multiplicity of selves sharing a common consciousness ... “If she who is now your mother had cohabited with someone else and had a son by him, and your father had done likewise, would you have come to be? Or were you living in them, and in your father’s father ... thousands of years ago? And even if this is so, why are you not your brother, why is your brother not you ... ?” To ponder why I am not my brother, or where or what I would be if my parents had not met, is to induce a kind of intellectual vertigo.
From *Nature* 2 October 1965

100 Years Ago

The Victoria Cross is given for conspicuous courage in the face of the enemy: and, if we want a good example of such courage, we have it in the English nurse, at the American Hospital in Neuilly, who inoculated herself with a pure culture of the bacilli of gas-gangrene, that she might help the discovery of the best treatment of that disease. We rejoice that she now is out of danger of death. The annals of medicine record many similar instances of self-devotion ... The protective treatments against cholera, plague, and typhoid fever were of course, well tested by their discoverers on themselves before they were put to national uses ... We salute with reverence this big-hearted Englishwoman, but we beg of her that she will not do it again.
From *Nature* 30 September 1915

A molecular tightrope

The identification of a regulatory site on the UBE3A protein that can be phosphorylated to alter its enzymatic activity provides insight into the aetiology of two human neurodevelopmental diseases, Angelman syndrome and autism.

YPE ELGERSMA

Fifty years ago, the paediatrician Harry Angelman discovered a genetic disorder of neuronal development¹ that was characterized by an unusually happy disposition, intellectual disability, absence of speech, impaired motor coordination, specific behavioural traits and treatment-resistant epilepsy^{1,2}. The disorder, now named Angelman syndrome, affects 1 in 20,000 children and is caused by genetic mutations that decrease the activity or expression of UBE3A, a ubiquitin ligase enzyme that targets proteins for degradation³. By contrast, duplications of the *UBE3A* gene cause another neurodevelopmental disorder, autism⁴, which is characterized by impaired social communication and a tendency to engage in repetitive behaviours. The mechanism by which UBE3A activity is regulated in the brain has been poorly understood. Writing in *Cell*, Yi *et al.*⁵ report that phosphorylation dynamically regulates this enzyme.

Humans have two copies of most genes, one inherited from each parent. Typically, cells express both copies equally. However, in the case of *UBE3A*, the paternally inherited gene is almost entirely silenced in neurons⁶, suggesting that *UBE3A* dosage is crucial for its function in these cells. Indeed, whereas mutations that reduce expression of the maternally inherited *UBE3A* gene are responsible for Angelman syndrome, duplication of the maternal-chromosome region in which *UBE3A* resides is associated with Dup15q syndrome, which involves developmental delay, autism, speech deficits and epilepsy^{7,8}. Having more than two copies of this maternal-chromosome region leads to an almost-certain chance that the individual will develop autism^{7,8}.

Because *UBE3A* dosage is crucial during neuronal development⁹, mechanisms must be in place to continually adjust neuronal UBE3A activity. Early insights into the regulation of UBE3A came from the discovery that the enzyme’s activity can be stimulated by cancer-causing viruses. For instance, the human papillomavirus E6 protein binds to and stimulates UBE3A, enhancing degradation of the p53 tumour-suppressor protein and so causing cervical cancer¹⁰. Several other cancer-causing viruses have independently evolved similar mechanisms for activating UBE3A and degrading tumour-suppressor

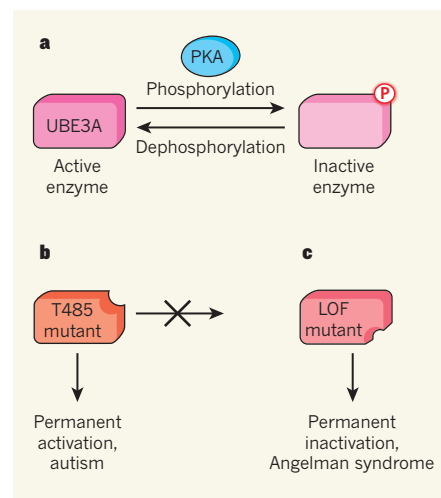


Figure 1 | Balancing act. Levels of activity of the enzyme UBE3A must be carefully regulated during neuronal development. **a**, Yi *et al.*⁵ report that the enzyme PKA inactivates UBE3A through phosphorylation at a threonine amino-acid residue, T485. In healthy cells, UBE3A toggles between inactive and active states to ensure the correct level of activity. **b**, Mutations that alter UBE3A activity can cause disease. The authors find that T485 mutations prevent phosphorylation, permanently activating UBE3A. This causes autism-like changes in neuronal development *in vitro*, and the mutation was found in a child with autism. **c**, By contrast, ‘loss-of-function’ (LOF) mutations inactivate UBE3A, leading to another neurodevelopmental disorder, Angelman syndrome.

proteins^{11,12}. More recently, binding of the human HERC2 protein to UBE3A has been identified¹³ as a mechanism by which the cell itself can activate UBE3A, but no mechanisms for regulating inhibition of the enzyme have previously been reported.

Yi and colleagues now reveal how neurons successfully balance along this molecular tightrope of UBE3A activity (Fig. 1). The authors’ investigations began when they noticed that a disproportionate number of mutations associated with Angelman syndrome were clustered in one region of the protein, suggesting that this region is involved in UBE3A function. Indeed, these mutations caused a loss of UBE3A activity. The authors then discovered that a threonine amino-acid residue (dubbed T485) in this region is an acceptor site for phosphates — groups that are covalently added to proteins by kinase enzymes to modulate the proteins’ activity. Specifically, T485 is

phosphorylated by protein kinase A (PKA), which is a well-known modulator of synaptic plasticity (changes in the strength of the synaptic connections between neurons) and cognitive function.

To test whether the addition of a phosphate group to T485 stimulates or inactivates UBE3A, the authors mutated the residue to prevent phosphorylation. This resulted in an overactive form of UBE3A. When neurons harbouring this mutation were grown *in vitro*, the development of synapses was profoundly altered. This is a notable finding, given the wealth of data linking synaptic dysfunction to autism. Moreover, Yi *et al.* identified the T485 mutation in a child with autism, highlighting the potential human relevance of this regulatory mechanism.

More genetic evidence is needed to precisely quantify the risk of autism associated with UBE3A phosphorylation at T485. Nonetheless, Yi and colleagues' findings provide a compelling model to explain the narrow range of tolerance for UBE3A activity during neuronal development, and the concomitant risk of Angelman syndrome and autism when tight control of the enzyme's activity is lost.

This model of UBE3A regulation opens up fresh avenues of investigation. The interplay between the PKA signalling pathway and UBE3A should now be fully explored. Of particular interest is the extent to which the role of PKA in synaptic plasticity and cognitive function is mediated through UBE3A phosphorylation.

Moreover, the contrasting roles of UBE3A in Angelman syndrome compared with autism and Dup15q syndrome should be investigated further. Identification of the downstream targets of UBE3A will help to reveal how changes in UBE3A activity can result in such distinct disorders. Additional human-genetics studies should be performed to determine whether other mutations in the PKA pathway are also associated with autism. And finally, it is important to determine whether stimulating PKA-dependent phosphorylation of UBE3A might be beneficial to people with autism. Drugs known as phosphodiesterase inhibitors are well-known stimulators of the PKA pathway, and could be used to test this.

Further investigation into the regulation and function of UBE3A has the potential to yield insights into the mechanisms that underlie a range of neurodevelopmental disorders, from epilepsy to intellectual disability and autism. This is encouraging progress, on the golden anniversary of the discovery of Angelman syndrome. ■

Ype Elgersma is at the Erasmus MC University Medical Center, ENCORE Expertise Center for Neurodevelopmental Disorders, Department of Neuroscience, 3015 CN Rotterdam, the Netherlands. e-mail: y.elgersma@erasmusmc.nl

1. Angelman, H. *Dev. Med. Child Neurol.* **7**, 681–688 (1965).
2. Williams, C. A. *Am. J. Med. Genet. C* **154C**, 432–437 (2010).
3. Kishino, T., Lalande, M. & Wagstaff, J. *Nature Genet.* **15**, 70–73 (1997).
4. Noor, A. *et al. Hum. Mutat.* **36**, 689–693 (2015).
5. Yi, J. J. *et al. Cell* **162**, 795–807 (2015).
6. Albrecht, U. *et al. Nature Genet.* **17**, 75–78 (1997).
7. Urraca, N. *et al. Autism Res.* **6**, 268–279 (2013).
8. Hogart, A., Wu, D., LaSalle, J. M. & Schanen, N. C. *Neurobiol. Dis.* **38**, 181–191 (2010).
9. Silva-Santos, S. *et al. J. Clin. Invest.* **125**, 2069–2076 (2015).
10. Scheffner, M., Huibregtse, J. M., Vierstra, R. D. & Howley, P. M. *Cell* **75**, 495–505 (1993).
11. Munakata, T. *et al. PLoS Pathog.* **3**, e139 (2007).
12. Louria-Hayon, I. *et al. Cell Death Differ.* **16**, 1156–1166 (2009).
13. Kühnle, S. *et al. J. Biol. Chem.* **286**, 19410–19416 (2011).

CONDENSED-MATTER PHYSICS

Flat transistor defies the limit

A transistor has been demonstrated that operates at low supply voltages by exceeding a theoretical limit. The finding opens up avenues to the development of integrated circuits that have extremely low power consumption. SEE LETTER p.91

KATSUHIRO TOMIOKA

Field-effect transistors (FETs) are used in integrated circuits that form part of commonplace devices such as smartphones, tablets and laptops. Improving the performance of these items depends crucially on miniaturizing FETs. But miniaturization cannot continue forever — it faces a fundamental thermionic limit below which a FET's turn-on performance (rapid switching on at low supply voltage) cannot improve without a commensurate increase in power consumption. On page 91 of this issue, Sarkar *et al.*¹ demonstrate a tunnel-FET (TFET) that combines an atomically thin, two-dimensional semiconducting crystal with a 3D germanium substrate and shows excellent turn-on performance at only 0.1 volts.

The miniaturization of FETs faces inherent problems caused by the effect of reducing the length of the channel (the layer through which current flows between the source and drain electrodes) and by large leakage currents when the transistor is turned off. These factors result in huge stand-by power consumption, but they can be suppressed in FETs that have a multi-gate structure, in which the whole surface of the FET's channel is covered with a metal electrode (the gate) and an electrically insulating material, together forming a gate stack. This enables good electrostatic control of the channel by the gate (see Fig. 1a of Sarkar and colleagues' paper¹).

The main outstanding goal for next-generation FETs is that they should offer both high-performance and low-power operation — after all, speed and long battery life are desirable qualities in electronic devices. For these demands to be satisfied, researchers must overcome two challenges. First, new

channel materials are needed. For example, metal-oxide-semiconductor (MOS) FETs are expected to move from the currently used 'strained' silicon-germanium chips to devices made from III–V compounds² (such as indium-gallium-arsenic), pure germanium³ or 2D semiconductors⁴, because such materials achieve high switch-on currents at low internal (gate-to-source) voltages. The second and harder challenge is to overcome the inability to scale down the supply voltage of FETs.

The switching properties of conventional FETs are dictated by a quantity known as the subthreshold slope (SS), which cannot be lower than the theoretical limit of 60 mV per decade of current at room temperature⁵; this means that a 60-mV increase in the voltage results in a 10-fold increase in the current. SS is inversely proportional to the rate at which the current that flows through the FET increases towards its 'on' value with increasing supply voltage. Hence, the supply voltage attainable for MOSFETs has a minimum value because the voltage is proportional to SS.

To overcome this physical limitation, researchers have investigated FETs that involve mechanisms such as tunnelling⁶, impact ionization⁷ and negative capacitance⁸, or that use mechanical switches⁹. Overall, tunnelling devices, such as TFETs, are promising because they can operate efficiently at low supply voltages and offer good compatibility with widely used complementary MOS (CMOS) technology. The operational principle of TFETs hinges on the transport of charge by means of a quantum-tunnelling mechanism: by using the gate to modulate the current, the SS is lowered below the classical theoretical limit. However, only a few TFETs have reported 'steep' SS values (less than the theoretical limit of 60 mV per decade of current over several current

decades). These are based on, for example, silicon- or germanium-based materials^{10,11}, silicon nanowires¹² and III–V–silicon interfaces (heterojunctions)¹³.

For TFEs to have steep SS values, devices must be designed with precise control of doping — the substitution of a small percentage of one type of atom for another, to change the amount of charge available to conduct current through the semiconductor. This property allows an internal electrical field to be effectively induced at the tunnelling junction. Also, established architectures such as the gate stacks of conventional MOSFETs need improvement.

Sarkar *et al.* propose a TFET made from a bilayer of molybdenum disulfide (MoS₂) and bulk germanium that overcomes many of the above challenges. The bilayer 2D MoS₂ crystal is placed on a germanium substrate, forming an ultra-thin junction through van der Waals bonding (see Fig. 1a of the paper¹). This custom-built heterojunction shows negative differential resistance (that is, the amount of current through the junction tends to decrease with increasing voltage across it), confirming the occurrence of tunnelling transport across the van der Waals bonds. Because it is not easy to make gate stacks from 2D MoS₂ materials, the authors use a solid polymer electrolyte as the device's gate electrode.

At room temperature, Sarkar and colleagues' TFET achieves a very steep SS (31.1 mV per decade of current, averaged over 4 decades); the minimum value reached is 3.9 mV per decade of current. This 2D semiconducting heterojunction ranks as one of the most promising materials for the fabrication of switches that could operate with supply voltages as low as 0.1 V, resulting in a reduction in power consumption of more than 90% compared with conventional FETs.

However, challenges remain. They include enhancing the robustness of the 2D crystal, controlling the properties of the gate's solid-state insulator, tuning the threshold voltage (above which the current increases nonlinearly with the supplied gate voltage), finessing the doped channel's structure and increasing the tunnelling current. Assuming that these issues can be addressed, Sarkar and co-workers' findings could lead to improved CMOS applications and, as the authors suggest, even to highly effective sensors for biological applications based on TFETs¹⁴. Junctions and materials such as those reported by the authors could contribute to the development of ultra-low-power and energy-efficient integrated circuits, which would find their way into mainstream electronic devices. ■

Katsuhiro Tomioka is in the Graduate School of Information Science and Technology, and the Research Center for Integrated Quantum Electronics (RCIQE), Hokkaido University, 060–0814 Sapporo, Japan.
e-mail: tomioka@rciqe.hokudai.ac.jp

1. Sarkar, D. *et al.* *Nature* **526**, 91–95 (2015).
2. del Alamo, J. A. *Nature* **479**, 317–323 (2011).
3. Pillarisetty, R. *Nature* **479**, 324–328 (2011).
4. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. *Nature Nanotechnol.* **6**, 147–150 (2011).
5. Ferrain, I., Collinge, C. A. & Colinge, J.-P. *Nature* **479**, 310–316 (2011).
6. Seabaugh, A. C. & Zhang, Q. *Proc. IEEE* **98**, 2095–2110 (2010).
7. Gopalakrishnan, K., Griffin, P. B. & Plummer, J. D. *IEEE Int. Electron Devices Meet.* 289–292 (2002).
8. Salahuddin, S. & Datta, S. *Nano Lett.* **8**, 405–410 (2008).
9. Pott, V. *et al.* *Proc. IEEE* **98**, 2076–2094 (2010).
10. Jeon, K. *et al.* *IEEE VLSI Technol. Symp.* 121–122 (2010).
11. Kim, S. H., Kam, H., Hu, C. & Liu, T.-J. K. *IEEE VLSI Technol. Symp.* 178–179 (2009).
12. Gandhi, R., Chen, Z., Singh, N., Banerjee, K. & Lee, S. *IEEE Electron Device Lett.* **32**, 437–439 (2011).
13. Tomioka, K., Yoshimura, M. & Takashi, F. *IEEE VLSI Technol. Symp.* 47–48 (2012).
14. Sarkar, D. & Banerjee, K. *Appl. Phys. Lett.* **100**, 143108 (2012).

HUMAN GENOMICS

The end of the start for population sequencing

In the final phase of a seven-year project, the genomes of 2,504 people across five continental regions have been sequenced. The result is a compendium of in-depth data on variation in human populations. SEE ARTICLES P.68 & P.75

EWAN BIRNEY & NICOLE SORANZO

The human genome comprises three billion bases, of which millions differ between any two genomes from different people. The 1000 Genomes Project, which was set up in 2008, aimed to study this variation in at least this many people, and in doing so to provide a solid foundation on which to build an understanding of genetic variation in the human population. Early work from the pilot phase of the project provided preliminary data sets and gave the first details about methods for analysing these data. In this issue of *Nature*, the 1000 Genomes Project publishes its final two papers^{1,2}, which analyse 2,504 genomes from 26 populations, and provide the most comprehensive view of global human variation so far.

In the first paper (page 68), the 1000 Genomes Project Consortium¹ focuses on relatively simple, short variations that affect up to 500 bases. As expected, the vast majority of variants affect only one base. However, the paper also outlines small but complex changes that were not studied in previous analyses. In the second paper (page 75), Sudmant *et al.*² explore more-complex changes that affect larger portions of the chromosome. These structural variants, which can be up to 500,000 bases long, are analysed in much greater detail than had been possible previously, thanks to improvements in genome sequencing over the past decade.

Sudmant *et al.* show that structural variants are abundant in human genomes and arise from a series of evolutionary processes that are more complex than had been thought³. The authors compare the effect of these variants on gene expression with the effect of variants

in which one base is substituted for another, known as single nucleotide polymorphisms (SNPs). They report that the structural variants have a disproportionate impact on gene expression, given their relatively low numbers in the genome compared with SNPs.

Both studies make huge strides in increasing the accuracy and sensitivity of DNA sequencing, particularly in identifying mutations that result in the insertion or deletion of bases (known as indels) and so shift the position of each subsequent base up or down the sequence. This advance allowed the consortium to correlate the presence of indels at specific positions with the more-humdrum SNPs.

These correlative data sets have benefits for genome-wide association studies (GWAS), which compare genomes from large cohorts of people to identify variants that are associated with disease or other traits. The correlated data sets from the 1000 Genomes Project Consortium¹ allow researchers to infer a large complement of variants, including indels and structural variants, in panels of people for whom only a small subset of SNPs have been analysed, using partial sequencing techniques such as genotyping arrays. Because genotyping arrays are cheap, the ability to infer variation allows researchers to focus on increasing sample sizes — a crucial next step in improving our understanding of the genetics of disease. Furthermore, panels inferred in this way should enable the identification of disease-associated variants that occur at substantially lower frequencies than can be identified by GWAS alone (the consortium identified variants that are present in as little as some 0.5% of the population with European ancestry, whereas these rare variants are found only sparsely in direct genotyping studies). The



Figure 1 | Genetic variation in human populations. Two studies^{1,2} have sequenced the genomes of 2,504 people from 26 populations across 5 continental regions — the broadest swathe of populations examined so far.

use of such panels will be particularly effective when combined with other whole-genome-sequencing data sets, such as that of the UK10K Consortium⁴, which is studying the genetic code of 10,000 people in fine detail.

Sudmant and colleagues also find that some structural variants occur in genomic regions that have been previously associated with complex traits or disease. The mechanisms by which these variants underpin disease risk remain largely unexplored, but the authors' data set provides the starting point for further mechanistic studies. Thus, the new data are expected to facilitate future exploration of rare structural variants that have previously been largely inaccessible.

Another major advance on previous phases of the project is the broad sampling of populations (Fig. 1). Sequences have been obtained from people in five continental regions (East and South Asia, Europe, Africa and the Americas). As expected, given the sub-Saharan origin of modern humans, the 1000 Genomes Project Consortium finds that most of the world's variation between humans occurs in sub-Saharan populations. The papers' repository of variants thus provides a much richer view than previous, Euro-centric data sets⁵ as to what constitutes normal variation in humans, and will enable cost-effective genetic studies in sub-Saharan populations. Indeed, such studies are already under way, designed to make use of data from the 1000 Genomes Project. Understanding how genetic variation can differ between people from different continents also affects our understanding both of recent human evolution and of medicine. The latter is particularly pertinent in cosmopolitan cities, where clinicians are increasingly assessing people from many ethnic backgrounds.

The papers also sequence admixed

populations, in which two previously separate populations have become mixed — for example, African American populations, which have African, European and Native American genetic heritage. Sequencing admixed populations is important because, for instance, it can help us to understand genetic variation in populations for whom few genomes have been sequenced, such as Native Americans. There are many admixed populations worldwide, including African American, Afro-Caribbean, Hispanic and North African populations, and the current studies lay the foundation for analysing and using genetic information from these groups.

The current data sets and analyses have been openly released and have already been used in thousands of publications.

Consistent with the aims of the 1000 Genomes Project, the current data sets and analyses have been openly released and have already been used in thousands of publications, ensuring that they will have a lasting impact. It is to the credit of the project that the people who donated DNA consented to the full release of their genetic data with the understanding that no other associated information, for example about health problems, would be collected. However, such completely accessible genome data sets are likely to become a minority, because there is a growing shift towards genomic data sets that are clinically annotated and so cannot normally be freely distributed. National laws and ethical standards mean that these data sets sometimes come with complex restrictions, even for use in research.

In this new world of genome sharing, baseline genetic data on human populations will

still be needed, and such data will be much more useful when openly released. The International Genome Sample Resource⁶, which was created earlier this year, provides a coordination centre for open data sets. However, given that controlled-access data are likely to become the norm, strategies that make it easier to use and reuse clinical data sets will be hugely beneficial. The nascent Global Alliance for Genomics and Health⁷ provides an international framework for discussion about these complex issues.

The future of human population genetics is both rosy, with many more data being produced, and complex, with more-involved ethics and more-strictly controlled access likely to be required. The 1000 Genomes Project has delivered the data and the methodological foundation for this future. ■

Ewan Birney is at the European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, UK.

Nicole Soranzo is at the Wellcome Trust Sanger Institute, Hinxton CB10 1HH, UK, and at the School of Clinical Medicine, University of Cambridge, UK.
e-mails: birney@ebi.ac.uk; ns6@sanger.ac.uk

1. The 1000 Genomes Project Consortium. *Nature* **526**, 68–74 (2015).
2. Sudmant, P. H. et al. *Nature* **526**, 75–81 (2015).
3. Alkan, C., Coe, B. P. & Eichler, E. E. *Nature Rev. Genet.* **12**, 363–376 (2011).
4. UK10K Consortium. *Nature* <http://dx.doi.org/10.1038/nature14962> (2015).
5. The 1000 Genomes Project Consortium. *Nature* **491**, 56–65 (2012).
6. www.internationalgenome.org
7. www.genomicsandhealth.org

E.B. declares competing financial interests. See go.nature.com/vy2wgw for details.

CRISPR–Cas immunity in prokaryotes

Luciano A. Marraffini¹

Prokaryotic organisms are threatened by a large array of viruses and have developed numerous defence strategies. Among these, only clustered, regularly interspaced short palindromic repeat (CRISPR)–Cas systems provide adaptive immunity against foreign elements. Upon viral injection, a small sequence of the viral genome, known as a spacer, is integrated into the CRISPR locus to immunize the host cell. Spacers are transcribed into small RNA guides that direct the cleavage of the viral DNA by Cas nucleases. Immunization through spacer acquisition enables a unique form of evolution whereby a population not only rapidly acquires resistance to its predators but also passes this resistance mechanism vertically to its progeny.

Prokaryotic viruses (bacteriophages or phages) are the most abundant life form, outnumbering their hosts by a factor of 10 (refs 1, 2). Since the beginning of the study of phages in the laboratory^{3,4}, investigators isolated and characterized bacteria that were resistant to phage infection^{5–7}. This led to the discovery of many different antiviral defence mechanisms (Box 1). CRISPR loci and CRISPR-associated (*cas*) genes encode a unique defence mechanism that provides rapid and robust adaptation to the rapidly evolving viruses (primarily double-stranded (ds)DNA viruses and other foreign DNA) of archaea and bacteria. CRISPR loci consist of an array of short (approximately 30–40 base pairs (bp)) and partially palindromic, repetitive sequences interspaced by equally short ‘spacer’ sequences (Fig. 1) from viral or plasmid origin. Spacers are at the centre of CRISPR defence as they specify immunity against phages or plasmids that contain a complementary sequence^{8,9}. The acquisition and utilization of spacer sequences constitute the two main stages of CRISPR immunity (Fig. 1). In the first stage, also known as ‘adaptation’ or ‘spacer acquisition’ (Fig. 1a), sequences from the viral genome are integrated into the CRISPR array (that is, the host is immunized). The second stage, where immunity is executed, can be further divided into two phases (Fig. 1b). First, in the guide RNA biogenesis phase, the CRISPR array is transcribed and processed to generate short RNAs containing one spacer sequence. Second, in the targeting phase, the spacer sequences in these RNAs are used as guides to direct the cleavage of the viral genome by the Cas endonucleases. As is the case with other major antiviral defence systems (Box 1), phages can escape CRISPR immunity through mutations in the target region that prevent its recognition and/or cleavage^{10,11}. However, in contrast to the other defence systems, the rapid acquisition of new spacers endows CRISPR–Cas loci with an in-built mechanism to counterattack these phage ‘escapers’^{10,12–14}. This distinctive property makes CRISPR–Cas immunity a unique form of heritable and adaptive immunity.

Discovery of CRISPR–Cas immune systems

CRISPR–*cas* loci and their function in antiviral and antiplasmid defence were first analysed *in silico*. The first description of CRISPR loci appeared in 1987, after the sequencing of the *iap* gene of *Escherichia coli*¹⁵. An ‘unusual’ repeat cluster was found downstream of this gene but due to the lack of homology to other known sequences the authors concluded that “the biological significance of these sequences is not known”. The first comprehensive report of CRISPR sequences was only possible after the accumulation of prokaryote genomes in GenBank¹⁶, and established CRISPR loci as common clusters of repetitive sequences

in bacteria and archaea. In 2002, the isolation and sequencing of small non-coding RNA species of *Archaeoglobus fulgidus* revealed that CRISPR loci are transcribed into small RNAs¹⁷. Also in 2002, the *cas* genes were identified as a gene family associated with CRISPR loci¹⁸. Sequence homology indicated that many of the Cas proteins participate in chemical reactions involving nucleic acids. In 2005 it was discovered that spacers match sequences present in phages and plasmids^{19–21}, and that the more spacers present in *Streptococcus thermophilus* strains, the fewer phages that were able to infect them¹⁹. These findings suggested a role for CRISPR–Cas systems in the prevention of phage infection and plasmid conjugation. All these data were incorporated into a model for CRISPR–Cas immunity reminiscent of RNA interference in eukaryotes, in which small CRISPR RNAs (crRNAs) had an antisense function against phage or plasmid transcripts, thereby preventing the propagation of these genetic elements^{20,22}, with the Cas proteins functioning as the effectors of the immunity mechanism²².

The initial experimental work on CRISPR–Cas systems concentrated on testing this model. Three foundational studies uncovered the basic features of CRISPR–Cas immunity: adaptation, the role of crRNA guides and the targeting of the invading nucleic acid (Fig. 1). The first study demonstrated the suspected function of CRISPR–Cas in the prevention of phage infection in *S. thermophilus*⁸. The work demonstrated a fundamental aspect of CRISPR–Cas immunity that was not foreseen in the models: that immunity is adaptive. To test for an involvement of CRISPR–*cas* loci in antiphage defence, Barrangou and colleagues infected *S. thermophilus* with phage and examined the CRISPR locus of the phage-resistant bacteria⁸. They observed that the resistant mutants harboured one or more new spacer sequences that perfectly matched a region of the genome of the infecting phage, which showed that new spacers (and immunity) are acquired during infection. This study also implicated the *cas* genes in CRISPR-mediated defence. A second study (Brouns and colleagues²³) demonstrated the requirement of crRNAs for immunity that was proposed in the early models. Brouns and colleagues discovered a Cas ribonucleoprotein complex harbouring the RNase responsible for the generation of small (mature) crRNAs in *E. coli*²³. Disruption of the complex or mutation of the catalytic residues of the RNase subunit prevented the accumulation of crRNAs and at the same time abrogated CRISPR immunity against phages. A third study demonstrated the role of CRISPR–Cas systems in the prevention of plasmid conjugation⁹. Marraffini and Sontheimer showed that a spacer present in the CRISPR locus of *Staphylococcus epidermidis*, which matched a region of the *nickase* gene of staphylococcal plasmids, prevented the conjugative transfer of such plasmids. The study also revealed

¹Laboratory of Bacteriology, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA.

BOX 1

Mechanisms of antiviral defence in prokaryotes

A common mechanism of defence against phage is the prevention of phage adsorption and/or genome injection. Bacteria can switch on or off the expression of phage receptors^{105,106} or secrete polysaccharides that limit the access to the receptor¹⁰⁷. Bacterial hosts can also express membrane proteins that prevent injection of the viral genetic material into the cell¹⁰⁸. Abortive infection is another common antiviral strategy in which the infected host cell sacrifices itself to prevent phage propagation¹⁰⁹. In most of the cases studied to date, phage infection is detected by a sensor protein that activates a cell death mechanism such as membrane depolarization¹¹⁰ or inhibition of bacterial translation¹¹¹. Other abortive infection mechanisms work through a phage-triggered activation of toxin-antitoxin systems that kills the host^{112,113}. Restriction-modification is the best studied defence system of prokaryotes and confers protection by cleaving the invading viral genome¹¹⁴. This system utilizes nucleases that recognize and cleave short DNA motifs. The target sequences of restriction nucleases are present in both the viral and host genome, but the bacterial chromosome is protected through methylation of the target DNA sequence. Therefore, restriction-modification systems display both endonuclease and methylase activities with the same sequence specificity. Other defence systems recently discovered, and less well characterized, are the bacteriophage exclusion (BREX) system, which blocks phage replication¹¹⁵, and the prokaryotic Argonaute pathway, which attacks foreign nucleic acid elements^{116,117}.

As expected from the highly dynamic virus-host interactions that occur in prokaryotic ecosystems, phages have evolved several counter-defence strategies to overcome host defences¹¹⁸. Phages can adapt to use different receptors¹¹⁹ or produce degrading enzymes to eliminate the extracellular polysaccharides that occlude receptors¹²⁰. Mutations in the viral proteins that trigger abortive infection allow phages to bypass this resistance mechanism¹⁰⁹, and when methylation of the invading phage DNA occurs faster than its cleavage by restriction endonucleases, the infecting phage can overcome this barrier and rapidly propagate throughout the whole host population¹¹⁴. To adapt to phage evasion strategies the host population needs to restore the lost mechanism of defence, in most cases through the selection of infrequent (and sometimes costly) mutations. By contrast, the acquisition of spacer sequences into CRISPR-Cas loci constitutes an in-built mechanism that provides a rapid and efficient response against phages that escape immunity through the introduction of mutations in the target site^{10,12-14}.

that as opposed to the proposed RNA-interference-like mechanism, CRISPR-Cas systems provide immunity by targeting DNA, rather than RNA. To show this, a self-splicing intron was placed in the *nickase* target sequence, creating a scenario in which the DNA target was interrupted but the RNA target was intact after splicing. Immunity to conjugation was lost, demonstrating the requirement of an intact DNA, but not RNA, target for CRISPR-Cas function. These results also suggested the existence of crRNA-guided, programmable Cas DNA nucleases, opening up the possibility for the development of CRISPR-based DNA manipulation tools²⁴.

Molecular mechanisms of CRISPR-Cas immunity

All CRISPR-Cas systems harbour the *cas1* and *cas2* genes, which are central to the immunization stage of the pathway (see below). However, based on the accessory *cas* gene content, each system can be classified into three different types²⁵. While all types use the same basic molecular mechanism to achieve immunity, that is, through crRNA-guided

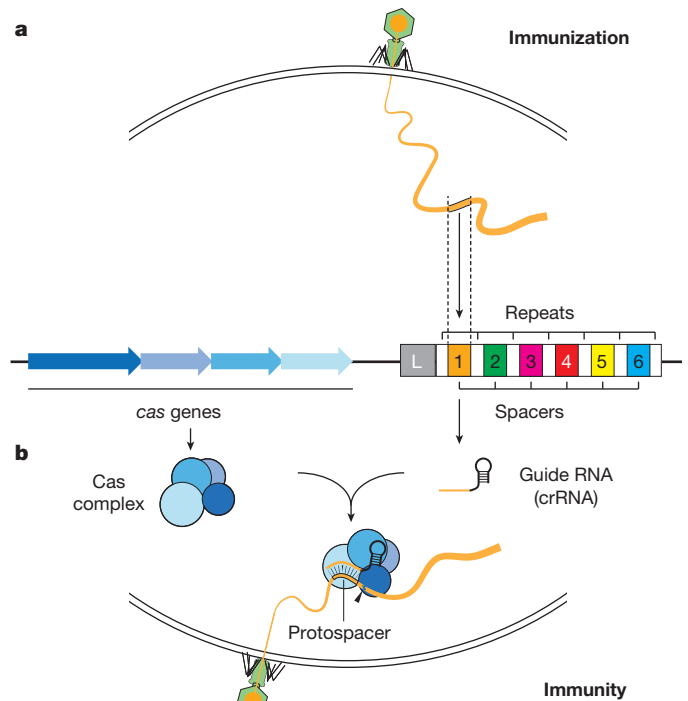


Figure 1 | Stages of CRISPR-Cas immunity. CRISPR loci are a cluster of short DNA repeats (white boxes) separated by equally short spacer sequences of phage and plasmid origin (coloured, numbered boxes). This repeat/spacer array is flanked by an operon of CRISPR-associated (*cas*) genes (blue-tone arrows) that encode the machinery for the immunization and immunity stages of the system. The CRISPR array is preceded by a leader sequence (grey box) containing the promoter for its expression. **a**, In the immunization stage, spacer sequences are captured upon entry of the foreign DNA into the cell and integrated into the first position of the CRISPR array. **b**, In the immunity stage the spacer is used to target invading DNA that carries a cognate sequence for destruction. Spacers are transcribed and processed into small CRISPR RNAs (crRNAs) in the 'crRNA biogenesis' phase. These small RNAs act as antisense guides for Cas RNA-guided nucleases (which usually form a complex) that locate and cleave the target sequence (black arrowhead) in the invader's genome during the 'targeting' phase.

nucleases, they differ in the biogenesis of crRNAs and the targeting requirements. Type I CRISPR-Cas immunity is mediated by the Cascade complex and the Cas3 nuclease (Fig. 2a)²³. One of the subunits of Cascade, Cas6, is a repeat-specific endoribonuclease that cleaves the precursor crRNA that is generated by transcription of the full CRISPR array^{23,26}. This cleavage produces short crRNAs that remain associated with Cascade and that are used by the complex to locate a complementary sequence in the target DNA^{23,27,28}, known as the protospacer. Another subunit, Cas8 (also known as CasA or Cse1), recognizes a short sequence motif located immediately upstream of the target sequence recognized by the crRNA²⁹. Sequence motifs adjacent to the targets specified by CRISPR spacers were first identified in type II systems^{11,19,30} and subsequently named as 'protospacer adjacent motif', or PAM³¹. PAM recognition is required for type I CRISPR-Cas immunity³², and the absence of a PAM in the repeat sequences prevents the targeting of the spacers within the CRISPR array by their complementary crRNAs; that is, it prevents an autoimmune reaction. The presence of a PAM promotes Cascade binding to its target^{29,33,34} and the formation of an R-loop between the crRNA spacer sequence and the dsDNA³⁵⁻³⁷. The first 8 bp at the 5' end of the crRNA-DNA duplex are critical for immunity and define a 'seed' sequence within the target^{28,32}. Mutant viruses containing mutations in this region can escape type I CRISPR immunity in *E. coli*³². An exception is mutations in the sixth nucleotide of the seed, which do not affect CRISPR immunity. The recent crystal structure of the interaction between a guide crRNA and its cognate

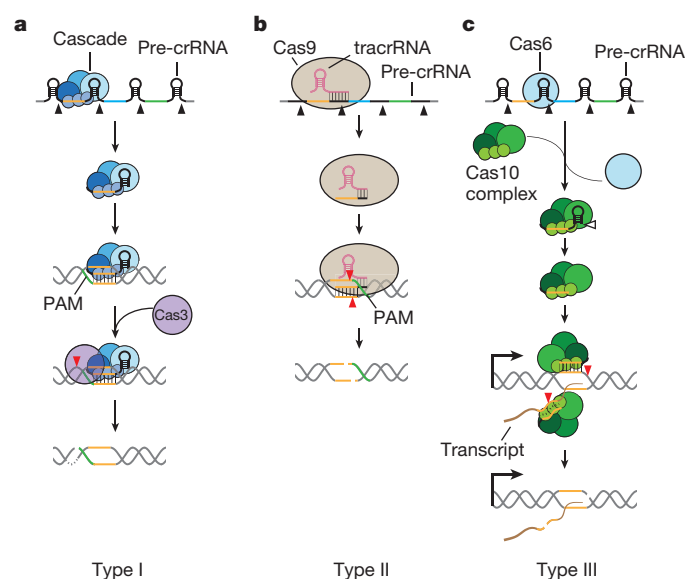


Figure 2 | Immunity mechanisms of the different CRISPR-Cas types.

a, Type I systems. A Cas protein complex known as Cascade cleaves at the base of the stem-loop structure of each repeat in the long precursor crRNA (pre-crRNA, black arrowheads), which generates short crRNA guides. The Cascade-crRNA complex scans the target DNA for a matching sequence (known as protospacer), which is flanked by a protospacer-adjacent motif (PAM, in green). Annealing of the crRNA to the target strand forms an R-loop; the Cas3 nuclease is recruited and cleaves the target downstream of the PAM (red arrowhead) and also degrades the opposite strand. **b, Type II systems.** These systems encode another small RNA known as trans-encoded crRNA (tracrRNA) which is bound by Cas9 and has regions of complementarity to the repeat sequences in the pre-crRNA. The repeat/tracrRNA dsRNA is cleaved by RNase III to generate crRNA guides for the Cas9 nuclease (black arrowheads). This nuclease cleaves both strands of the protospacer/crRNA R-loop (red arrowhead). A PAM (in green) is located downstream of the target sequence. **c, Type III systems.** Cas6 is a repeat-specific endoribonuclease that cleaves the pre-crRNA at the base of the stem-loop structure of each repeat (black arrowhead). The crRNA is loaded into the Cas10 complex where it is further trimmed at the 3' end to generate a mature crRNA (white arrowhead). The Cas10 complex requires target transcription to cleave the non-template strand of the protospacer DNA and it is also capable of crRNA-guided transcript cleavage (red arrowheads).

single-stranded (ss)DNA target within the Cascade complex showed that the crRNA-ssDNA interaction forms a non-canonical ribbon structure in which every sixth nucleotide is rotated out of the double helix and therefore not engaged in the formation of a base pair^{38–40}. Finally, target recognition by Cascade triggers the recruitment and activity of Cas3, a nuclease that introduces ssDNA breaks into the target virus or plasmid^{33,41–44} to initiate their degradation.

Type II CRISPR-Cas systems require only one *cas* gene, *cas9*, to execute immunity in the presence of an existing targeting spacer sequence (Fig. 2b)⁴⁵. However, as opposed to the other CRISPR types, two small RNAs are needed for immunity: the crRNA and the trans-encoded crRNA (tracrRNA)⁴⁶. The tracrRNA forms a secondary structure that mediates its association with Cas9 (refs 47–49) but also has a region that is complementary to the repeat sequences of the CRISPR array⁴⁶. The dsRNA formed between the tracrRNA and the precursor crRNA is cleaved by RNase III, resulting in the cleavage of each repeat and the processing of the long CRISPR transcript into small crRNA guides⁴⁶. Type II immunity also requires a PAM, and mutations in this motif are the most common mechanism of escape from CRISPR immunity by the targeted viruses¹¹. In contrast to type I, the PAM is located immediately downstream of the target sequence^{11,19} and is recognized by a PAM-binding domain present in Cas9 (refs 47–49). Type II CRISPR-Cas immunity results in the introduction of crRNA-specific dsDNA breaks in the invading DNA⁵⁰ that require two nuclease

domains: HNH and RuvC⁴⁵. Each of these domains cleaves one DNA strand of the protospacer sequence^{51,52} and the tracrRNA is absolutely required for cleavage⁵². The first step in target recognition is the transient binding of Cas9 to PAM sequences within the target DNA, which promotes the melting of the two DNA strands immediately upstream of the PAM⁵³. A productive interaction in this region of the target, between 6–8 bases of the spacer sequence of the crRNA guide and the melted DNA (the 'seed' sequence of type II systems^{11,54}), triggers the formation of an R-loop and target cleavage^{37,53}.

In type III CRISPR-Cas systems, the precursor crRNA is cleaved by a repeat-specific endoribonuclease, Cas6, which is not part of a complex⁵⁵. As a result of this processing, 8 nucleotides of the repeat sequence remain at the 5' end of the spacer sequence in the crRNA^{55,56}, a sequence known as the crRNA tag. By an as yet unknown mechanism, the small crRNAs generated after Cas6 cleavage are transferred to a larger complex⁵⁷, the Cas10–Csm or Cas10–Cmr complex for type III-A or III-B systems, respectively. Within these complexes the crRNAs undergo a process of maturation whereby the 3' end is trimmed at 6-nucleotide intervals^{58–60}. As opposed to type I and II systems, in which targeting relies strictly on the recognition of DNA sequences, type III CRISPR-Cas immunity also requires target transcription^{61,62} and a crRNA complementary to the non-template strand of the DNA target and to the transcript⁶². Both DNA^{9,61–63} and RNA^{63–68} are targeted by type III CRISPR-Cas systems, resulting in the co-transcriptional crRNA-guided cleavage of the target DNA and its transcripts^{63,69}. The Cas10 complex contains both nucleolytic activities: the palm domain of Cas10 is required for cleavage of the non-template DNA strand⁶³, and backbone subunits Csm3 (refs 63, 65) or Cmr4 (ref. 66), for type III-A or III-B systems, respectively, are responsible for cleavage of the RNA transcripts. To date, no PAM requirements have been observed for type III CRISPR-Cas targeting. To avoid targeting of the CRISPR locus, type III systems rely on the differential base pairing between the crRNA tag and the sequences flanking the protospacer⁵⁶. Whereas the absence of complete complementarity between these sequences licenses DNA targeting, full complementarity between the crRNA tag and the repeat sequence in the CRISPR locus prevents DNA targeting^{56,63,67} and thus autoimmunity. The biological significance of this elaborate targeting mechanism is beginning to be elucidated. The transcription requirement for targeting offers the possibility of immunological tolerance of mobile genetic elements with non-transcribed regions, such as prophages with the potential to provide a fitness advantage to the host⁶². Moreover, when the prophage re-activates and enters the lytic cycle, its genome is transcribed and type III targeting resumes, resulting in clearance of the infection and preventing the death of the host cell. Studies have shown that RNA cleavage can protect against RNA viruses⁶⁶; however, the role of RNA targeting in immunity against DNA elements and possibly regulation of gene expression remains to be determined.

CRISPR immunization

The mechanisms by which new spacer sequences are added to the CRISPR locus to immunize the host are beginning to be understood. This process can be divided into two stages: the selection of protospacer sequences from the invader DNA and their integration into the CRISPR array (Fig. 3). With a few exceptions, the spacer acquisition mechanism has been studied in detail in the *E. coli* type I CRISPR-Cas system. This is because an early report showed that overexpression of type I Cas1 and Cas2 is sufficient for the expansion of the *E. coli* CRISPR array of this organism and thus established a very simple and elegant experimental system to study this phenomenon⁷⁰.

Central to the mechanism of selection of new spacer sequences is the prevention of autoimmunity; that is, the ability of the acquisition machinery to distinguish self (chromosomal) from non-self (invading) DNA. Failure to do so leads to the death of the host, a scenario that has been tested and exploited for the use of CRISPR-Cas systems as genome editing tools of bacteria^{54,71–73} and as antimicrobials^{74–76}. Applying the Cas1–Cas2 overexpression system—in which autoimmunity is avoided

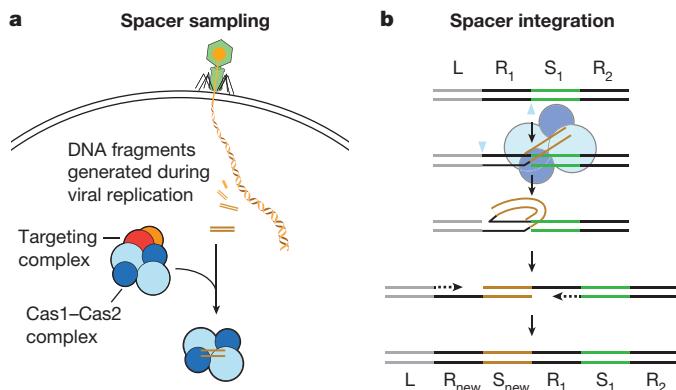


Figure 3 | Mechanism of CRISPR immunization. **a**, The first step of CRISPR immunization is the sampling of the spacer sequences. These are believed to be generated from non-specific DNA breaks that occur during replication of the virus or plasmid. The fragments generated are captured by the Cas1–Cas2 complex, with the participation of the targeting machinery for the recognition of DNA sequences carrying a functional PAM. **b**, The Cas1–Cas2 complex catalyzes the integration of the spacer into the first position of the CRISPR array. Cas1 performs two concerted cleavage–ligation reactions whereby the 5' end of each repeat strand is cleaved (blue arrowhead) and ligated to the 3' ends of the spacer. This mechanism generates two ssDNA gaps on the repeat sequences that flank the inserted spacer, which presumably are filled by DNA polymerase (dotted arrow). L, leader; R₁, first repeat; R₂, second repeat; R_{new}, new repeat; S₁, first spacer; S_{new}, new spacer.

by using an *E. coli* strain that lacks the Cascade and Cas3 immunity machinery—several studies determined that there is a strong preference for the integration of plasmid over chromosomal spacer sequences^{70,77,78}. Recently, it has been shown that Chi sites, 8-nucleotide motifs present approximately once every 5 kb in the *E. coli* genome, limit the acquisition of chromosomal sequences⁷⁹. Chi sites are over-represented in the *E. coli* chromosome⁸⁰; therefore, this mechanism favours the acquisition of spacer sequences from foreign elements. The origin of the substrates for CRISPR adaptation is not yet known. In *E. coli*, the overexpression of Cas1–Cas2 in the absence of immunity leads to the acquisition of spacer sequences derived from regions limited by Chi sites on one side and by *Ter* sites on the other⁷⁹. There are two *Ter* sites in the *E. coli* genome, *TerA* and *TerC*, which are diametrically opposed from the origin of replication to stall a faster replication fork and ensure proper chromosomal decatenation at the end of a round of DNA replication⁸¹. Replisome stalling at either *Ter* site leads to chromosome breaks that are processed by the RecBCD exonuclease complex, generating 3' OH overhang ends for RecA-mediated homologous recombination repair of the DNA lesion⁸². To avoid excessive DNA degradation, RecBCD activity is inhibited by Chi sequences⁸³. Therefore, the chromosomal acquisition data obtained in *E. coli* (ref. 79) suggest that the nucleolytic processing of dsDNA breaks generated during replication is the source of new spacer sequences for the CRISPR–Cas acquisition machinery. This model is supported by a reduction in spacer acquisition from the host chromosome when DNA replication is chemically inhibited and in mutants that lack the RecBCD machinery⁷⁹. Although not directly tested, this work predicts that the DNA degradation that follows the formation of spontaneous breaks during viral or plasmid replication produces DNA fragments that are captured by the Cas adaptation machinery for their incorporation as non-self spacers into the CRISPR array (Fig. 3a).

A second aspect of spacer selection relates to the PAM requirement for targeting: how does the acquisition machinery ensure that a new spacer will match a protospacer flanked by the correct PAM? Recently it was revealed that Cas9, which has the ability to recognize the PAM, is used to select new spacers with the correct PAM sequence⁸⁴ (Fig. 3a). Genetic studies showed that Cas9 is absolutely required for type II CRISPR–Cas adaptation^{84,85}. While the nuclease activity does not con-

tribute to the role of Cas9 in spacer acquisition^{84,85}, mutation of the PAM-binding domain results in the incorporation of new spacers that match DNA targets but without conservation of their flanking sequences; that is, without a PAM⁸⁴. Supporting a role for Cas9 in the acquisition of functional spacer sequences it was shown that this nuclease forms a complex with Cas1, Cas2 and Csn2, proteins exclusively involved in CRISPR adaptation, and that swapping *cas9* alleles that recognize different PAM sequences results in the incorporation of spacers with PAMs corresponding to the *cas9* allele used⁸⁴.

Cas1 and Cas2, without Cascade, are sufficient for *E. coli* type I spacer acquisition. However, it is not clear whether the new spacers acquired in the absence of Cascade display an absolute conservation of the PAM sequence—AAG for this CRISPR–Cas system^{70,86,87}. Interestingly, the first (5' end) nucleotide of *E. coli* spacer sequences is invariably G; that is, the last nucleotide of the PAM^{88–90}. Therefore it is possible that, similarly to Cas9 in type II systems, the PAM recognition feature of the Cascade complex is involved in the selection of functional spacers. Although this remains to be tested, studies have found another role for Cascade in type I adaptation, in a phenomenon known as 'priming'⁸⁶. It has been shown that the presence of pre-existing (priming) spacers with partial homology (and therefore unable to provide full immunity) to the ssDNA phage M13 increases the rate of spacer acquisition by several orders of magnitude⁸⁶. In addition, new spacers acquired in the presence of a priming spacer have a strong strand bias, producing a crRNA guide that matches the same strand matched by the priming spacer^{86,87}. Primed acquisition facilitates adaptation against invaders that are related in sequence to previous invaders (that have partial matches to pre-existing spacers) and also against escapers (that is, phage containing target mutations that prevent CRISPR–Cas immunity). The mechanism of primed acquisition is yet to be resolved. Primed adaptation requires the Cascade complex⁸⁶, and the partial match between the crRNA–target sequences that trigger this process results in a distinct mode of Cascade binding to the target DNA³⁵. It is believed that this binding results in a low incidence of target cleavage that, similarly to the DNA breaks that occur at replication stall sites⁷⁹, would generate the substrates for spacer acquisition. Interestingly, primed acquisition reveals a coupling between the immunity and immunization stages of CRISPR defence.

In the second stage of CRISPR immunization, after the selection of the spacer by the acquisition machinery, spacers are integrated into the CRISPR array in a reaction that resembles retroviral integration (Fig. 3b). Recently, it was shown that *E. coli* Cas1 and Cas2 perform the spacer integration reaction. Studies using the Cas1–Cas2 overexpression system demonstrated the presence of integration intermediates *in vivo*⁹¹. Using specific probes for Southern blot assays it was demonstrated that each strand of the first repeat sequence is separated and ligated to the 3' end of the new spacer sequence. This reaction is catalysed by the Cas1–Cas2 complex, which is composed of two Cas1–Cas2 units, and mutations that prevent the interaction between these proteins⁷⁸ or eliminate Cas1 nuclease activity^{70,78,91} abrogate spacer acquisition. The complex binds to *E. coli* CRISPR sequences *in vitro*⁷⁸ and, when incubated with a 33-nucleotide dsDNA substrate (which mimics a captured spacer sequence), it mediates the covalent addition of this dsDNA into plasmids harbouring the CRISPR array⁹². This reaction requires free 3' OH ends on the 33-nucleotide spacer substrate, which presumably perform a direct nucleophilic attack on the phosphodiester bond between the first repeat and spacer sequences⁹² (Fig. 3b). Importantly, deep sequencing of the reaction products revealed a strong bias for the integration of spacers with a C nucleotide at the 3' end (that is, the complementary base to the first (5' end) G of the spacer sequence⁹²). Collectively, these results indicate that the Cas1–Cas2 complex provides the orientation specificity of the integration reaction, which probably occurs by two consecutive 3' OH attacks on opposite sides of the repeat sequence, with the first nucleophilic attack by the 3'-end C of the spacer sequence on the 5' end of the bottom strand of the repeat. After the repair of the repeat gaps that result from this reaction, a new repeat-spacer unit is added into the CRISPR array. This mechanism

is consistent with early work that demonstrated that the additional repeat is derived from the first repeat of the CRISPR array⁷⁰. This study showed that a 'labelling' mutation introduced in the first, but not the second, repeat sequence is incorporated into the new repeat after spacer acquisition.

Perspectives

An extensive body of work has established the function of CRISPR-Cas systems as the adaptive immune system of prokaryotes, which function in protection against viral infection and plasmid invasion. One intriguing aspect is the role of these systems in the regulation of horizontal gene transfer (HGT) between microorganisms. HGT is fundamental for the generation of the genetic diversity required for prokaryotic evolution and is achieved through three major routes: phage transduction, plasmid conjugation and DNA transformation⁹³. In addition to the activity of CRISPR-Cas against phages, it has been experimentally demonstrated that CRISPR-Cas systems can prevent plasmid conjugation⁹ and the transformation of naturally competent bacteria^{94,95}. Therefore, in principle, CRISPR immunity can prevent the major routes of HGT and therefore have an important role in limiting the evolution of bacteria and archaea. This drawback of CRISPR-Cas systems has often been considered a cause for the inconsistent distribution of these loci^{96,97}, which are present in only ~50% of bacterial and ~90% of archaeal genomes⁹⁸. A recent study investigated the relationship between the number of spacer sequences (an indicator of CRISPR activity) and the estimated number of HGT transfer events for all the genomes present in GenBank⁹⁹. If CRISPR immunity limits HGT, a negative correlation between these two values is expected. Since this correlation was not evident, the study suggests that there is no impact of CRISPR-Cas on gene transfer among prokaryotes over evolutionary timescales. Future work should determine the effects of CRISPR-Cas immunity on HGT at the population level as well as other potential disadvantages of the system, such as a high incidence of autoimmunity and/or off-target effects, which could impact on its distribution among prokaryotic organisms.

Another puzzling aspect of CRISPR-Cas systems is their high diversity^{25,100}. For each of the three CRISPR types there are many subtypes encoding divergent sets of Cas proteins. The biological significance of this diversity is not known. A driving force for the diversification of CRISPR-Cas systems could be imposed by phage-encoded anti-CRISPR mechanisms. It has been shown that phages encode small proteins that specifically inhibit type I-F CRISPR immunity¹⁰¹, which could conceivably lead to the evolution of another type or subtype to overcome the inhibition of the existing CRISPR-Cas system. In this hypothetical scenario CRISPR-Cas systems would diversify as a result of the arms race with the viruses they target. It is also possible that the mechanistic differences between types and subtypes provide advantages for different CRISPR-Cas systems in different conditions imposed by the lifestyle or the environment of the host. An example of this is the type III systems, which tolerate untranscribed or inert mobile genetic elements⁶². Different conditions where this tolerance is beneficial or detrimental will determine the selection of type III systems in different hosts. Similarly, mechanistic differences between other types or subtypes that have an impact on the fitness of the host could have a role in the distribution of the CRISPR systems. Furthermore, additional functions of particular CRISPR subtypes could explain their diversification. For example, the type II-C CRISPR-Cas system of the bacterial pathogen *Francisella novicida* harbours a tracrRNA with sequence homology to an endogenous lipoprotein gene^{102,103}. When complexed with Cas9, the tracrRNA associates with another small RNA, the small CRISPR-associated RNA (scaRNA), to mediate the silencing of the lipoprotein gene during *F. novicida* infection. Another example is the RNA cleavage feature of type III (Cas10) complexes, which offers the possibility of regulation of gene expression by these systems^{67,104} that could lead to the preferential selection of these systems over other types that cannot perform this function.

Similarly to restriction-modification systems, the possibilities to manipulate DNA sequences in a predictable manner by CRISPR-Cas immunity have led to numerous applications in molecular biology, most notably the genetic engineering of a range of different cell types²⁴. This highlights the importance of the study of basic biological problems, particularly the interactions between prokaryotes and the mobile genetic elements that invade them, for the development of innovative biotechnological applications. In this context it is exciting to consider the different mechanisms of immunity present in the understudied (or even presently undiscovered) CRISPR-Cas systems, as well as the many uncharacterized prokaryotic defence systems carried by the archaeal and bacterial domains of life that inhabit every corner of our planet.

Received 29 May 2015; accepted 7 August 2015.

1. Bergh, O., Borsheim, K. Y., Bratbak, G. & Haldal, M. High abundance of viruses found in aquatic environments. *Nature* **340**, 467–468 (1989).
2. Chibani-Chennoufi, S., Bruttin, A., Dillmann, M. L. & Brussow, H. Phage-host interaction: an ecological perspective. *J. Bacteriol.* **186**, 3677–3686 (2004).
3. d'Herelle, F. Sur un microbe invisible antagoniste des bacilles dysentériques. *C.R. Acad. Sci. Paris* **165**, 373–375 (1917).
4. Twort, F. W. An investigation on the nature of ultra-microscopic viruses. *Lancet* **186**, 1241–1243 (1915).
5. Burnet, F. M. Further observations on the nature of bacterial resistance to bacteriophage. *J. Pathol. Bacteriol.* **32**, 349–354 (1929).
6. Gratia, A. Studies on the D'herelle phenomenon. *J. Exp. Med.* **34**, 115–126 (1921).
7. Luria, S. E. & Delbruck, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511 (1943).
8. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
9. **A study that demonstrated that CRISPR-Cas loci provide acquired immunity against bacteriophages.** Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845 (2008).
10. **A paper showing that CRISPR-Cas loci target DNA molecules in a sequence-specific manner, highlighting for the first time the potential for the technological applications of these systems.** Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**, 1047–1050 (2008).
11. **This study revealed the arms race between CRISPR-Cas systems and viruses in their natural habitat.** Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).
12. Childs, L. M., England, W. E., Young, M. J., Weitz, J. S. & Whitaker, R. J. CRISPR-induced distributed immunity in microbial populations. *PLoS ONE* **9**, e101710 (2014).
13. Paez-Espino, D. *et al.* CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *MBio* **6**, e00262–15 (2015).
14. Weinberger, A. D. *et al.* Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput. Biol.* **8**, e1002475 (2012).
15. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.* **169**, 5429–5433 (1987).
16. Mojica, F. J., Diez-Villasenor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.* **36**, 244–246 (2000).
17. **First description of CRISPR loci as a new family of repetitive sequences in prokaryotes.** Tang, T. H. *et al.* Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl Acad. Sci. USA* **99**, 7536–7541 (2002).
18. Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).
19. **First description of cas sequences as a family of genes associated with CRISPR repeats.** Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindromic repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
20. **This paper, along with references 20 and 21, made the discovery that spacer sequences match viruses and plasmids, and suggested a defence function for CRISPR-Cas systems.** Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).
21. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
22. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational

- analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7 (2006). **This work provided the first comprehensive model for the mechanism of CRISPR-Cas immunity.**
23. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008). **Study demonstrating the central role of crRNA guides and Cas ribonucleoproteins in CRISPR immunity.**
 24. Pennisi, E. The CRISPR craze. *Science* **341**, 833–836 (2013).
 25. Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nature Rev. Microbiol.* **9**, 467–477 (2011).
 26. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).
 27. Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Struct. Mol. Biol.* **18**, 529–536 (2011).
 28. Wiedenheft, B. *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl Acad. Sci. USA* **108**, 10092–10097 (2011).
 29. Sashital, D. G., Wiedenheft, B. & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol. Cell* **46**, 606–615 (2012).
 30. Horvath, P. *et al.* Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1401–1412 (2008).
 31. Mojica, F. J., Díez-Villasenor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
 32. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl Acad. Sci. USA* **108**, 10098–10103 (2011).
 33. Sinkunas, T. *et al.* *In vitro* reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. *EMBO J.* **32**, 385–394 (2013).
 34. Westra, E. R. *et al.* CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol. Cell* **46**, 595–605 (2012).
 35. Blosser, T. R. *et al.* Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Mol. Cell* **58**, 60–70 (2015).
 36. Rutkauskas, M. *et al.* Directional R-loop formation by the CRISPR-Cas surveillance complex cascade provides efficient off-target site rejection. *Cell Rep.* <http://dx.doi.org/10.1016/j.celrep.2015.01.067> (2015).
 37. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl Acad. Sci. USA* **111**, 9798–9803 (2014).
 38. Jackson, R. N. *et al.* Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* **345**, 1473–1479 (2014).
 39. Mulepati, S., Heroux, A. & Bailey, S. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484 (2014).
 40. Zhao, H. *et al.* Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. *Nature* **515**, 147–150 (2014).
 41. Hochstrasser, M. L. *et al.* CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl Acad. Sci. USA* **111**, 6618–6623 (2014).
 42. Huo, Y. *et al.* Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nature Struct. Mol. Biol.* **21**, 771–777 (2014).
 43. Mulepati, S. & Bailey, S. *In vitro* reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J. Biol. Chem.* **288**, 22184–22192 (2013).
 44. Sinkunas, T. *et al.* Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30**, 1335–1342 (2011).
 45. Sapranas, R. *et al.* The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* **39**, 9275–9282 (2011).
 46. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
 47. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
 48. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).
 49. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
 50. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010). **This work showed that the crRNA-guided DNA targeting by CRISPR-Cas systems results in sequence-specific DNA cleavage.**
 51. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl Acad. Sci. USA* **109**, E2579–E2586 (2012).
 52. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
 53. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
 54. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature Biotechnol.* **31**, 233–239 (2013).
 55. Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22**, 3489–3496 (2008).
 56. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568–571 (2010).
 57. Sokolowski, R. D., Graham, S. & White, M. F. Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system. *Nucleic Acids Res.* **42**, 6532–6541 (2014).
 58. Hale, C., Kleppe, K., Terns, R. M. & Terns, M. P. Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* **14**, 2572–2579 (2008).
 59. Hatoum-Aslan, A., Maniv, I. & Marraffini, L. A. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc. Natl Acad. Sci. USA* **108**, 21218–21222 (2011).
 60. Hatoum-Aslan, A., Samai, P., Maniv, I., Jiang, W. & Marraffini, L. A. A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *J. Biol. Chem.* **288**, 27888–27897 (2013).
 61. Deng, L., Garrett, R. A., Shah, S. A., Peng, X. & She, Q. A novel interference mechanism by a type IIIB CRISPR-Cmr module in *Sulfolobus*. *Mol. Microbiol.* **87**, 1088–1099 (2013).
 62. Goldberg, G. W., Jiang, W., Bikard, D. & Marraffini, L. A. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature* **514**, 633–637 (2014).
 63. Samai, P. *et al.* Co-transcriptional DNA and RNA cleavage during type III CRISPR-Cas immunity. *Cell* **161**, 1164–1174 (2015).
 64. Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956 (2009). **First demonstration that some CRISPR-Cas systems can cleave RNA molecules.**
 65. Staals, R. H. *et al.* RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. *Mol. Cell* **56**, 518–530 (2014).
 66. Tamulaitis, G. *et al.* Programmable RNA shredding by the type III-A CRISPR-Cas system of *Streptococcus thermophilus*. *Mol. Cell* **56**, 506–517 (2014).
 67. Zebec, Z., Manica, A., Zhang, J., White, M. F. & Schleper, C. CRISPR-mediated targeted mRNA degradation in the archaeon *Sulfolobus solfataricus*. *Nucleic Acids Res.* **42**, 5280–5288 (2014).
 68. Zhang, J. *et al.* Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol. Cell* **45**, 303–313 (2012).
 69. Peng, W., Feng, M., Feng, X., Liang, Y. X. & She, Q. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Res.* **43**, 406–417 (2014).
 70. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
 71. Oh, J. H. & van Pijkeren, J. P. CRISPR-Cas9-assisted recombineering in *Lactobacillus reuteri*. *Nucleic Acids Res.* **42**, e131 (2014).
 72. Jiang, Y. *et al.* Multigene editing in the *Escherichia coli* genome via the CRISPR-Cas9 system. *Appl. Environ. Microbiol.* **81**, 2506–2514 (2015).
 73. Huang, H., Zheng, G., Jiang, W., Hu, H. & Lu, Y. One-step high-efficiency CRISPR/Cas9-mediated genome editing in *Streptomyces*. *Acta Biochim. Biophys. Sin.* **47**, 231–243 (2015).
 74. Bikard, D. *et al.* Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nature Biotechnol.* **32**, 1146–1150 (2014).
 75. Citorik, R. J., Mimee, M. & Lu, T. K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nature Biotechnol.* **32**, 1141–1145 (2014).
 76. Goma, A. A. *et al.* Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *MBio* **5**, e00928–e00913 (2014).
 77. Díez-Villasenor, C., Guzmán, N. M., Almendros, C., García-Martínez, J. & Mojica, F. J. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol.* **10**, 792–802 (2013).
 78. Nuñez, J. K. *et al.* Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nature Struct. Mol. Biol.* **21**, 528–534 (2014).
 79. Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 (2015). **This paper showed that dsDNA breaks generated during replication trigger spacer acquisition.**
 80. El Karoui, M., Biauudet, V., Schbath, S. & Gruss, A. Characteristics of Chi distribution on different bacterial genomes. *Res. Microbiol.* **150**, 579–587 (1999).
 81. Neylon, C., Kralicek, A. V., Hill, T. M. & Dixon, N. E. Replication termination in *Escherichia coli*: structure and antihelicase activity of the Tus-Ter complex. *Microbiol. Mol. Biol. Rev.* **69**, 501–526 (2005).
 82. Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **72**, 642–671 (2008).
 83. Smith, G. R. How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. *Microbiol. Mol. Biol. Rev.* **76**, 217–228 (2012).
 84. Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199–202 (2015).
 85. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev.* **29**, 356–361 (2015).
 86. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).

This study showed that pre-existing spacers with partial homology to an invader sequence enhance the acquisition of new spacers.

87. Swarts, D. C., Mosterd, C., van Passel, M. W. & Brouns, S. J. CRISPR interference directs strand specific spacer acquisition. *PLoS ONE* **7**, e35888 (2012).
88. Goren, M. G., Yosef, I., Auster, O. & Qimron, U. Experimental definition of a clustered regularly interspaced short palindromic duplication in *Escherichia coli*. *J. Mol. Biol.* **423**, 14–16 (2012).
89. Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. & Severinov, K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol.* **10**, 716–725 (2013).
90. Shmakov, S. *et al.* Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res.* **42**, 5907–5916 (2014).
91. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res.* **42**, 7884–7893 (2014).
92. Nuñez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193–198 (2015).
- This study showed the molecular mechanism of spacer integration.**
93. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Rev. Microbiol.* **3**, 711–721 (2005).
94. Bikard, D., Hatoum-Aslan, A., Mucida, D. & Marraffini, L. A. CRISPR interference can prevent natural transformation and virulence acquisition during *in vivo* bacterial infection. *Cell Host Microbe* **12**, 177–186 (2012).
95. Zhang, Y. *et al.* Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Mol. Cell* **50**, 488–503 (2013).
96. Jiang, W. *et al.* Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. *PLoS Genet.* **9**, e1003844 (2013).
97. Marraffini, L. A. CRISPR-Cas immunity against phages: its effects on the evolution and survival of bacterial pathogens. *PLoS Pathog.* **9**, e1003765 (2013).
98. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172 (2007).
99. Gophna, U. *et al.* No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. *ISME J.* **9**, 2021–2027 (2015).
100. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. The basic building blocks and evolution of CRISPR-cas systems. *Biochem. Soc. Trans.* **41**, 1392–1400 (2013).
101. Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493**, 429–432 (2013).
102. Sampson, T. R., Saroj, S. D., Llewellyn, A. C., Tzeng, Y. L. & Weiss, D. S. A. CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497**, 254–257 (2013).
103. Sampson, T. R., Saroj, S. D., Llewellyn, A. C., Tzeng, Y. L. & Weiss, D. S. Corrigendum: A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **501**, 262 (2013).
104. Hale, C. R. *et al.* Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell* **45**, 292–302 (2012).
105. Liu, M. *et al.* Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science* **295**, 2091–2094 (2002).
106. Zaleski, P., Wojciechowski, M. & Piekarczyk, A. The role of Dam methylation in phase variation of *Haemophilus influenzae* genes involved in defence against phage infection. *Microbiology* **151**, 3361–3369 (2005).
107. Hanlon, G. W., Denyer, S. P., Olliff, C. J. & Ibrahim, L. J. Reduction in exopolysaccharide viscosity as an aid to bacteriophage penetration through *Pseudomonas aeruginosa* biofilms. *Appl. Environ. Microbiol.* **67**, 2746–2753 (2001).
108. Lu, M. J. & Henning, U. Superinfection exclusion by T-even-type coliphages. *Trends Microbiol.* **2**, 137–139 (1994).
109. Molineux, I. J. Host-parasite interactions: recent developments in the genetics of abortive phage infections. *New Biol.* **3**, 230–236 (1991).
110. Parma, D. H. *et al.* The Rex system of bacteriophage lambda: tolerance and altruistic cell death. *Genes Dev.* **6**, 497–510 (1992).
111. Bingham, R., Ekunwe, S. I., Falk, S., Snyder, L. & Kleanthous, C. The major head protein of bacteriophage T4 binds specifically to elongation factor Tu. *J. Biol. Chem.* **275**, 23219–23226 (2000).
112. Aizenman, E., Engelberg-Kulka, H. & Glaser, G. An *Escherichia coli* chromosomal “addiction module” regulated by guanosine 3',5'-bispyrophosphate: a model for programmed bacterial cell death. *Proc. Natl Acad. Sci. USA* **93**, 6059–6063 (1996).
113. Fineran, P. C. *et al.* The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc. Natl Acad. Sci. USA* **106**, 894–899 (2009).
114. Bickle, T. A. & Kruger, D. H. Biology of DNA restriction. *Microbiol. Rev.* **57**, 434–450 (1993).
115. Goldfarb, T. *et al.* BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183 (2015).
116. Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D. K. & Aravin, A. A. Bacterial argonaute samples the transcriptome to identify foreign DNA. *Mol. Cell* **51**, 594–605 (2013).
117. Swarts, D. C. *et al.* DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* **507**, 258–261 (2014).
118. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nature Rev. Microbiol.* **8**, 317–327 (2010).
119. Doulatov, S. *et al.* Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* **431**, 476–481 (2004).
120. Sutherland, I. W., Hughes, K. A., Skillman, L. C. & Tait, K. The interaction of phage and biofilms. *FEMS Microbiol. Lett.* **232**, 1–6 (2004).

Acknowledgements L.A.M. is supported by the Rita Allen Scholars Program, an Irma T. Hirsch Award, a Sinsheimer Foundation Award and a NIH Director's New Innovator Award (1DP2AI104556-01).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence should be addressed to L.A.M. (marraffini@rockefeller.edu).

The origins of high hardening and low ductility in magnesium

Zhaoxuan Wu^{1,2} & W. A. Curtin¹

Magnesium is a lightweight structural metal but it exhibits low ductility—connected with unusual, mechanistically unexplained, dislocation and plasticity phenomena—which makes it difficult to form and use in energy-saving lightweight structures. We employ long-time molecular dynamics simulations utilizing a density-functional-theory-validated interatomic potential, and reveal the fundamental origins of the previously unexplained phenomena. Here we show that the key $\langle c + a \rangle$ dislocation (where $\langle c + a \rangle$ indicates the magnitude and direction of slip) is metastable on easy-glide pyramidal II planes; we find that it undergoes a thermally activated, stress-dependent transition to one of three lower-energy, basal-dissociated immobile dislocation structures, which cannot contribute to plastic straining and that serve as strong obstacles to the motion of all other dislocations. This transition is intrinsic to magnesium, driven by reduction in dislocation energy and predicted to occur at very high frequency at room temperature, thus eliminating all major dislocation slip systems able to contribute to c -axis strain and leading to the high hardening and low ductility of magnesium. Enhanced ductility can thus be achieved by increasing the time and temperature at which the transition from the easy-glide metastable dislocation to the immobile basal-dissociated structures occurs. Our results provide the underlying insights needed to guide the design of ductile magnesium alloys.

Developing lightweight structural metal is a crucial step on the path towards reduced energy consumption in many industries, especially automotive^{1,2} and aerospace³. Magnesium (Mg) is a lightweight metal, with a density that is 23% that of steel and 66% that of aluminium, and so has tremendous potential to achieve energy efficiency⁴. In spite of this tantalizing property, Mg generally exhibits low ductility, insufficient for the forming and performance of structural components. The low ductility is associated with the inability of hexagonal-close-packed (hcp) Mg to deform plastically in the crystallographic $\langle c \rangle$ direction, which is accomplished primarily by dislocation glide on the pyramidal II plane with the $\langle c + a \rangle$ Burgers vector⁵ (see Fig. 1). Experiments reveal a range of unusual, confounding, conflicting, and mechanistically unexplained phenomena connected to $\langle c + a \rangle$ dislocations that coincide with the inability of Mg to achieve high plastic strains⁶. Uncovering and controlling the fundamental behaviour of $\langle c + a \rangle$ dislocations is thus the key issue in using Mg, and any solution would catapult Mg science, technology, and applications forward. Success would enable, for instance, lightweight automobiles that would consume less energy, independent of the energy source, and thus act as a multiplier for many other energy-reduction strategies.

Because of its critical importance and promise, $\langle c + a \rangle$ slip has been extensively studied over five decades^{7–10}, and reports on this system are being published at an increasing rate^{11–14}. In Mg single crystals, $\langle c + a \rangle$ slip occurs predominantly on the pyramidal II system, and measurements show that pyramidal II $\langle c + a \rangle$ dislocations can glide at low stresses of ~ 20 MPa at low temperatures¹⁵ (77–133 K). However, transmission electron microscopy (TEM) studies frequently find $\langle c + a \rangle$ dislocations lying, mysteriously, along basal planes and coexisting with $\langle c \rangle$ and $\langle a \rangle$ dislocations^{9,16–19}. Under c -axis compression, single crystal Mg also exhibits a rapid increase in stress with increasing strain, that is, high work hardening^{9,15,18,19}, and fractures at low strains at temperatures up to 500 K (refs 9, 15, 19). New work shows the formation of a very high density of $\langle c + a \rangle$ dislocation loops also dissociated on, and

lying on, the basal planes¹⁴, with similar observations dating back fifty years^{7,8,20}. Many mechanisms have been proposed, all of which invoke extrinsic effects such as ‘heating’ by the electron beam⁸, vacancy and self-interstitial precipitation²¹, and dislocation obstacles of ‘unknown’ nature⁷. Also surprisingly, when loaded at slow strain rates ($\sim 10^{-1}$ – 10^{-5} s^{−1}), and particularly under compression, the yield strengths of both single crystal Mg^{8,15–17} and polycrystal Mg alloys^{22,23} increase with increasing temperature, that is, they show an anomalous temperature dependence, within a certain temperature range. This behaviour is hypothesized as being due to decomposition of the $\langle c + a \rangle$ dislocation into $\langle c \rangle$ and $\langle a \rangle$ (ref. 17), which is supported by TEM studies^{16,21} where junction pairs of $\langle c \rangle$, $\langle a \rangle$, and $\langle c + a \rangle$ dislocations were found. However, no decomposition process has been directly observed. It has also been proposed that the junctions may be formed by reactions of $\langle c \rangle$ and $\langle a \rangle$ dislocations²⁴. The $\langle c \rangle$ dislocation is also unusual because it too lies on the basal plane^{8,12,25}. Overall, how $\langle c + a \rangle$ and $\langle c \rangle$ dislocations can come to lie on the basal plane is the subject of discussion. Nor is there understanding of why the behaviour of $\langle c + a \rangle$ dislocations varies widely over a range of loading conditions, single-crystal orientations, polycrystalline texture, and temperature. Here, we provide an explanation for all of these unusual observed phenomena that is entirely intrinsic to hcp Mg.

Mechanisms of $\langle c + a \rangle$ dislocation transition

We first show the transitions of the pyramidal II $\langle c + a \rangle$ edge dislocation using molecular dynamics (MD) simulations (see Supplementary Information I and II for details). Properties of the screw dislocation (Extended Data Figs 1 and 7) are not directly relevant. We start the simulations with an initial $\langle c + a \rangle$ edge dislocation dissociated into $\frac{1}{2}\langle c + a \rangle + \frac{1}{2}\langle c + a \rangle$ on the pyramidal II glide plane, a structure in good agreement with density functional theory (DFT) calculations²⁶. We then execute long-time MD at elevated temperatures (500 K, 600 K, 700 K) to accelerate any thermally activated processes, and under

¹Institute of Mechanical Engineering, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. ²Institute of High Performance Computing, 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Singapore.

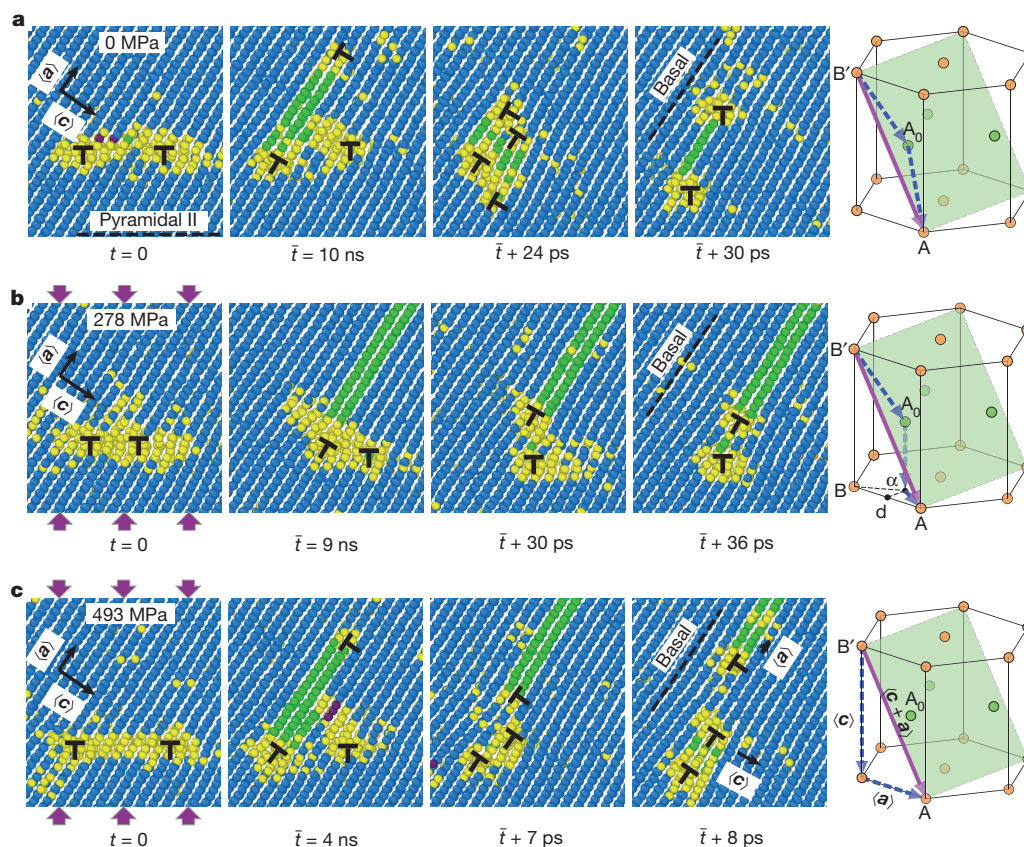


Figure 1 | Transitions of the easy-glide pyramidal II edge $\langle c+a \rangle$ dislocation. The $\langle c+a \rangle$ dislocation transforms into basal-dissociated products, as observed during long-time MD simulations, for **a**, zero, **b**, moderate, and **c**, high compressive stresses (indicated by the purple arrows) normal to the pyramidal II plane (green plane in rightmost images). The first four columns show MD simulations at the indicated times; the rightmost column shows the transition of dislocation Burgers vectors in the hcp unit cell. All cases start from the same dislocation (leftmost image) and show a distinct thermally activated intermediate state (second image) before undergoing a rapid intrinsic ‘climb-like’ dissociation onto the basal plane (third and fourth images). The final core structure (fourth image) depends on the applied load, with increasing applied load able to drive an $\langle a \rangle$ or a partial $\langle a \rangle$ dislocation

away from the $\langle c+a \rangle$ core. \bar{t} is the mean transition time to reach the intermediate state shown. Dislocation cores are indicated by the symbol ‘L’. Atoms in the atomistic images in this figure (first four columns) and subsequent ones are coloured on the basis of common neighbour analysis³⁷: blue, hcp; green, face-centred cubic (fcc); purple, body-centred cubic (bcc); yellow, all others. Dislocation core atoms thus appear predominantly as yellow. In **a–c**, the rightmost column shows the composition of the $\langle c+a \rangle$ Burgers vector before (solid purple arrow) and after (dashed blue arrows) the transition; the green and orange circles depict the two alternating layers of atoms in basal planes. In the rightmost column in **b**, d is the midpoint between A and B , and α is the vertical projection of A_0 onto the basal plane.

various applied stresses normal to the glide plane. As shown in Fig. 1, the initial $\langle c+a \rangle$ core is metastable and transforms into one of three new basal-oriented dislocations. The transition into the final state is always preceded by an intermediate state where a partial $\langle a \rangle$ (Shockley) dislocation is nucleated on the basal plane and glides away with a trailing basal I_2 intrinsic stacking fault⁵. Since the time to achieve this intermediate state (indicated in Fig. 1) is much longer than the subsequent transition to any of the final states, the transition into this intermediate state is the kinetically limiting step in the overall process. An applied normal stress exerts a resolved shear stress on the basal plane that determines the glide of the partial $\langle a \rangle$, which in turn determines the final structure. The frequencies of occurrence of the various final structures as a function of load are shown in Extended Data Fig. 2.

Figure 1a shows the transition that predominates at zero and low stresses. The newly nucleated partial $\langle a \rangle$ stays close to the original nucleation site and another partial $\langle a \rangle$ is then nucleated from the other half $\langle c+a \rangle$ dislocation. The two remaining partials then ‘climb’ in opposite directions to form a new $\langle c+a \rangle$ core dissociated on the basal plane and separated by a basal I_1 stacking fault⁵. The ‘climb’-like transition is atomistically complicated, involving significant atomic motions within the core region, but does not involve any vacancy diffusion or interstitial diffusion from the surrounding bulk because

the simulation contains no vacancies or interstitials. The core spreading on the basal plane is driven by the repelling force between the two partials, but the spreading is kinetically difficult because further climb would presumably require some sort of vacancy/interstitial pair formation and transport (see Extended Data Figs 3 and 4 and Supplementary Information III and IV). This final dislocation core has been observed previously^{8,11} and has recently been resolved in high-angle annular dark field scanning TEM (HAADF STEM)¹⁴. Figure 2a shows the atomic structure obtained from MD quenched to 0 K. In Fig. 2b we show a superposition of the atomic images from MD (projected into the plane of view perpendicular to the dislocation line) and from HAADF STEM (which shows spots due to diffraction from aligned columns of atoms) for one of the two basal-oriented $\langle c+a \rangle$ partials with the I_1 stacking fault emerging from it. Good agreement can be seen outside the core region. In the core region, the MD results show complex variations in atomic position along the dislocation line; this is consistent with the absence of clear spots in the HAADF STEM image that suggests an absence of structural order.

Figure 1b shows the transition that occurs mainly at intermediate loads. In this case, the first nucleated leading partial $\langle a \rangle$ glides away, leaving behind a wide I_2 stacking fault and a $\frac{1}{2}\langle c \rangle + d\alpha$ dislocation (see rightmost panel in Fig. 1b for nomenclature). The other half $\langle c+a \rangle$ on the pyramidal II plane then reacts with the residual dislocation to

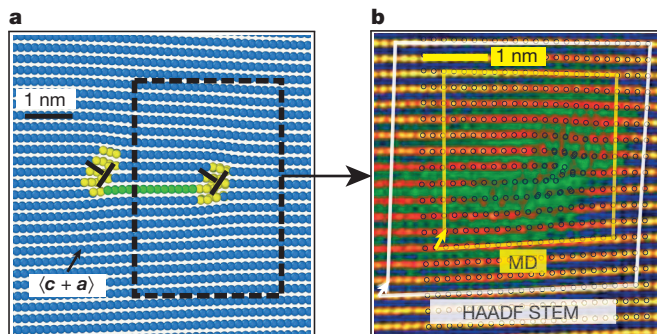


Figure 2 | Comparison of the $\langle c + a \rangle$ dislocation core structure in MD and experiments. **a**, Structure of a $\langle c + a \rangle$ dislocation climb-dissociated on the basal plane, from MD simulations quenched to 0 K. **b**, Superposition of the atomistic structure of one $\langle \frac{1}{2}c + p \rangle$ core (p is the Burgers vector of the Shockley partial $\langle a \rangle$ dislocation) as computed by MD (open circles), and the HAADF STEM image of the same core¹⁴, for which bright spots indicate well-defined columns of atoms through the thickness of the experimental specimen; a Burgers loop indicating the $\langle \frac{1}{2}c + p \rangle$ Burgers vector is shown for both images. HAADF STEM image from ref. 14: The structure of $\langle c + a \rangle$ type dislocation loops in magnesium. J. Geng *et al. Philosophical Magazine Letters*, 3 June 2014, reprinted by permission of the publisher (Taylor & Francis Ltd, <http://www.tandfonline.com>).

form another new ‘climb-dissociated’ dislocation with Burgers vector $\frac{1}{2}\langle c \rangle + B'A_0$. If the applied stress is then reduced, this core transforms into the previous core because the partial $\langle a \rangle$ is pulled back to the main dislocation. Figure 1c shows the transition that predominates at higher loads. In this case, the pyramidal II $\langle c + a \rangle$ decomposes into $\langle c \rangle$ and $\langle a \rangle$ dislocations by nucleating a trailing partial $\langle a \rangle$ behind the initial leading partial $\langle a \rangle$. The full $\langle a \rangle$ then glides away, driven by the high resolved shear stress, leaving behind a $\langle c \rangle$ dislocation. The $\langle c \rangle$ then also ‘climb-dissociates’ along the basal plane into two partial $\langle c \rangle$ dislocations separated by a basal extrinsic stacking fault, consistent with TEM observations²⁵ and HAADF STEM¹². Details of all dislocation reactions are shown in Extended Data Fig. 5 and in Supplementary Information V.

Dislocation transition rate and energy barrier

The $\langle c + a \rangle$ transitions are thermally activated processes. Figure 3a shows the measured mean transition time \bar{t} versus stress and temperature for this random process. \bar{t} depends on temperature and, weakly, on stress. Simulation cell sizes and boundary conditions have insignificant effects on the measured transition time (see Extended Data Fig. 6 and Supplementary Information VI and VII). The mean transition rate $R = 1/\bar{t}$ can be related to the normal-stress-dependent transition energy barrier $\Delta E(\sigma_{\text{norm}})$ using the Arrhenius law $R = \nu_0 \exp(-\Delta E(\sigma_{\text{norm}})/(kT))$, where ν_0 , k , and T are the attempt frequency, Boltzmann constant, and temperature, respectively. Estimating $\nu_0 = 10^{13} \text{ s}^{-1}$, the energy bar-

rier versus applied stress is shown in Fig. 3b. Overall, the energy barrier is $\sim 0.5 \text{ eV}$. This yields a fast transition rate of $\sim 10^5 \text{ s}^{-1}$ at 300 K and a very slow rate of $\sim 10^{-4} \text{ s}^{-1}$ at 150 K, consistent with the observation of anomalous strengthening only above $\sim 150 \text{ K}$ (refs 16, 17). Preliminary nudged elastic band^{27,28} calculations (Supplementary Information VIII) of the 0 K activation barrier yield an energy barrier of $\sim 0.6 \text{ eV}$, in good agreement with the estimates in Fig. 3. The applied stress has a weak asymmetric effect on nucleation because it influences the barrier by exerting a stress that moves the first partial $\langle a \rangle$ dislocation away (compressive) or towards (tensile) the original $\langle c + a \rangle$.

Dislocation energy

The observed transitions are driven by a reduction in total dislocation energy: all the new dislocations have lower energy than the easy-glide pyramidal II $\langle c + a \rangle$. The total dislocation energy per unit length within a cylindrical region of radius r centred at the dislocation core can be written as $E_{\text{tot}} = E_{\text{struc}} + K \ln(r/r_{\text{min}})$ for $r > r_{\text{min}}$ (see Extended Data Fig. 7 and Supplementary Information IX). Here, E_{struc} is the dislocation energy within a minimum radius r_{min} around the core region that contains all of the energy associated with the specific structure, such as the core energy, stacking fault energy, interactions among these structures, and near-field elastic energy. Conversely, the second term $K \ln(r/r_{\text{min}})$ is the additional elastic energy between r_{min} and r from the core region, which captures the elastic energy outside of the core region. The constant K is completely determined by the anisotropic elastic constants, Burgers vector b , and dislocation line direction. Here, all the $\langle c + a \rangle$ dislocations and total products have the same K , Burgers vector $b = c + a$, and line direction. Therefore, the total energy difference between any two structures is computed directly by the difference in their total energies E_{tot} measured at large distances $r > r_{\text{min}}$ from the complex core region⁵. Figure 4a shows E_{tot} versus $\ln(r/r_{\text{min}})$ for $r_{\text{min}} = 6b = 6|c + a|$ calculated atomistically for the three localized edge $\langle c + a \rangle$ type dislocations found here (see Fig. 4b). All three curves are parallel, as expected for $E_{\text{tot}}(r > r_{\text{min}})$, and the energy differences are precisely the differences in E_{struc} between the different structures. The easy-glide $\langle c + a \rangle$ dislocation on the pyramidal II plane has the largest energy per unit length, and the $\langle c + a \rangle$ dissociated on the basal plane has the smallest energy, $0.3 \text{ eV } \text{\AA}^{-1}$ lower, and is the most stable at zero applied stress. The $\langle c \rangle + \langle a \rangle$ remaining in close proximity has an intermediate energy, and is thus also stable relative to the easy-glide structure. In fact, while not shown, a related calculation can be performed to demonstrate that even the case of $\langle c \rangle$ and $\langle a \rangle$ separated to infinity has a lower total energy than does the easy-glide core, although higher than does the $\langle c \rangle + \langle a \rangle$ in close proximity, indicating that short-range interactions between the $\langle c \rangle$ and $\langle a \rangle$ dislocations reduce the overall energy relative to well-separated ones. There is thus an energy barrier for driving the $\langle a \rangle$ dislocation away from the $\langle c \rangle$.

The relative order of the edge dislocation energies rationalizes the simulations and various experimental observations. Under zero or

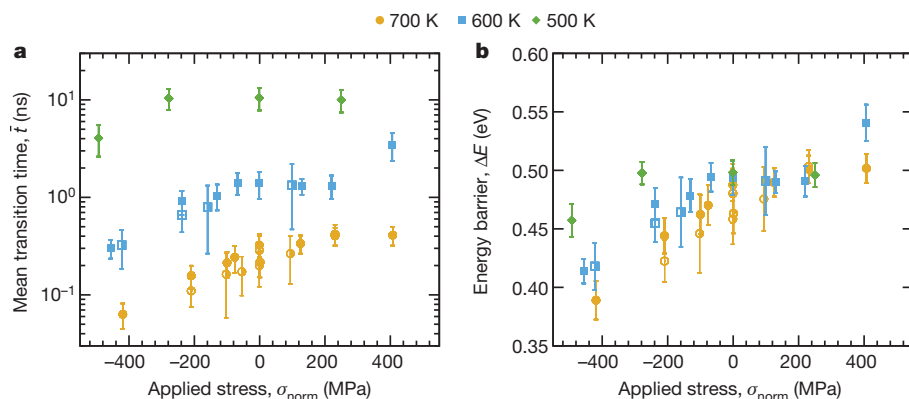


Figure 3 | Thermally activated mean transition time and energy barrier for the pyramidal II to basal plane transformation. Data are shown for three temperatures, see key. **a**, Mean transition time \bar{t} versus applied stress σ_{norm} as measured in MD for the dissociation events shown in Fig. 1. **b**, Energy barrier ΔE for the thermally activated transitions, showing a small dependence on temperature and applied stress. Error bars (s.e.m., $n = 2$) indicate the 95% confidence intervals of the mean transition time and energy barrier. Open, half-filled and quarter-filled symbols indicate (nearly identical) results obtained for larger simulation cells and different boundary conditions (see Supplementary Information VI).

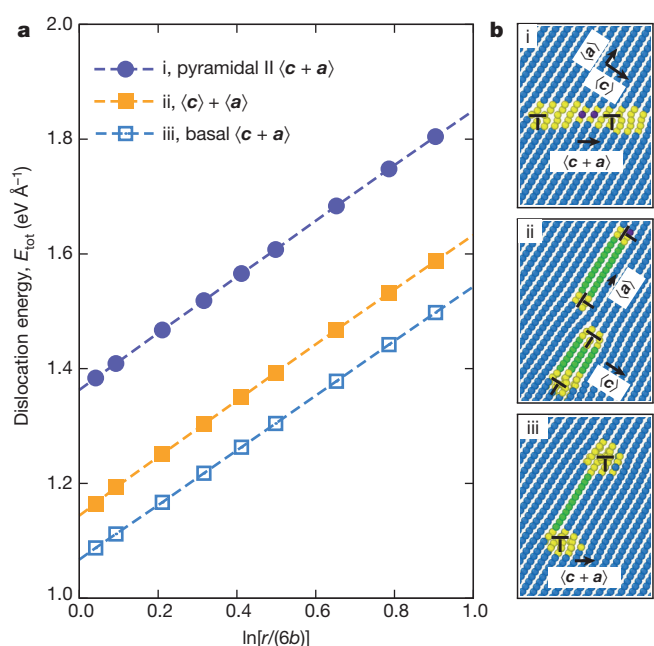


Figure 4 | Dislocation energy versus dislocation structure. **a**, Total energy within a cylindrical region of radius $r > r_{\min} = 6b$ for the $\langle c+a \rangle$ edge dislocation on the pyramidal II plane (purple circles; case i), for edge $\langle c \rangle$ and $\langle a \rangle$ in close proximity (orange squares; case ii), and for $\langle c+a \rangle$ edge dislocation climb-dissociated on the basal plane (open blue squares; case iii). Here, b is the magnitude of the $\langle c+a \rangle$ dislocation Burgers vector. The differences in energies between the three different dislocation structures are equal to the constant differences between the energies versus r , as expected by elasticity theory. **b**, Dislocation core structures corresponding to the energies shown in **a** (cases i, ii, and iii) as computed at zero temperature.

low applied normal stresses, the preferred transition is the lowest energy ‘climb-dissociated’ $\langle c+a \rangle$ on the basal plane. This is consistent with experimental observations in single crystal c -axis compression tests, where the resolved shear stress on basal planes is negligible and a high density of $\langle c+a \rangle$ loops/dislocations dissociated on basal planes^{14,19} or aligned with basal planes⁹ are observed. When there is a resolved stress τ on the basal plane, there is energy gained by the work

done by the applied field in moving the full/partial $\langle a \rangle$ dislocation away from the $\langle c \rangle$. When the full or partial $\langle a \rangle$ has moved a distance l , the total energy reduction per unit length is $\Delta E = \tau al$ for the full $\langle a \rangle$ and $\Delta E = (\tau al/2) - \gamma_{I_2} l$ for the partial $\langle a \rangle$, where γ_{I_2} is the I_2 stacking fault energy (~ 20 – 30 mJ m⁻²), and $a = |\mathbf{a}|$. These two cores can thus have the lowest total energy at sufficiently large τ . This is consistent with experimental observations in polycrystal Mg (ref. 21), where resolved shear stresses exist on basal planes in most grains and $\langle c \rangle$ and $\langle a \rangle$ dislocations/junctions are more commonly seen as compared to $\langle c+a \rangle$ basal loops.

Critical resolved shear stresses for dislocation glide

The three transformed cores shown in Fig. 1 are essentially immobile, as expected owing to their ‘climb dissociation’ onto the (non-glide) basal planes. Figure 5 shows the motion of the various dislocations under an applied resolved shear stress on the pyramidal II plane (Fig. 5a–d) and prism plane (Fig. 5e) at 300 K, as obtained via MD simulations (see Supplementary Information X). The initial $\langle c+a \rangle$ on the pyramidal II plane glides at stresses of ~ 11 MPa. All of the basal-dissociated dislocations are immobile up to stresses of more than 30 times the pyramidal II glide stress (see Extended Data Fig. 8 for the non-Schmid effect); only the $\langle a \rangle$ can glide away at ~ 119 MPa, as discussed above, leaving the immobile $\langle c \rangle$. A stress of ~ 430 MPa is needed to recombine the $\langle a \rangle$ and $\langle c \rangle$ into the $\langle c+a \rangle$, indicating a high energy barrier for this reverse reaction, so that this proposed mechanism for creation of $\langle c+a \rangle$ is unlikely²⁴.

Origin of low ductility and high strain hardening

The time-dependent thermal activation of the easy-glide pyramidal II $\langle c+a \rangle$ to immobile lower-energy, basal-dissociated $\langle c+a \rangle$ and $\langle c \rangle$ dislocations explains the low ductility of Mg. Generalized plasticity requires the operation of at least five independent slip systems⁵. Mg cannot sustain dislocation slip in the crystallographic $\langle c \rangle$ direction because the necessary $\langle c+a \rangle$ and $\langle c \rangle$ dislocations transform to immobile dislocations. Thus, only twinning remains to provide some deformation in the $\langle c \rangle$ direction, and the plastic strain due to twinning is very small (at most, $\sim 7\%$). Grains in a polycrystal can thus mainly slip only in the $\langle a \rangle$ direction, and grains oriented favourably for $\langle c+a \rangle$ slip are effectively rigid. High constraint stresses rapidly develop, leading to the early onset of fracture and low ductility.

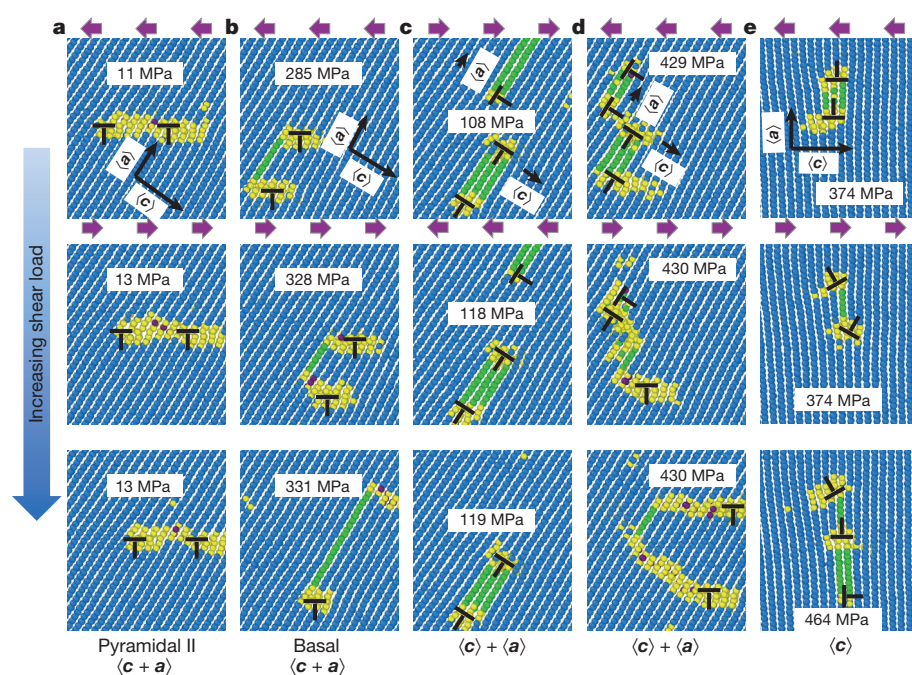


Figure 5 | Glide behaviour of the various dislocations under applied resolved shear stresses (directions indicated) at 300 K. Applied shear stress (indicated in each image) increases from top to bottom. **a**, Easy-glide pyramidal II $\langle c+a \rangle$, with glide starting at ~ 11 MPa. **b**, Basal-dissociated $\langle c+a \rangle$, with glide starting at a very high stress, ~ 330 MPa. **c**, $\langle a \rangle$ dislocation glides away from the remaining $\langle c+a \rangle$ product at ~ 119 MPa, leaving an immobile $\langle c \rangle$. **d**, Reaction of $\langle a \rangle$ and $\langle c \rangle$, forming the basal $\langle c+a \rangle$ dislocation for a resolved shear of ~ 430 MPa in the reverse direction. **e**, Absence of glide for the basal-dissociated $\langle c \rangle$ up to ~ 464 MPa, where nucleation of a partial $\langle a \rangle$ dislocation occurs.

Low ductility is also driven by high hardening rates, as measured in Mg, particularly for single crystals or textured polycrystals oriented for $\langle c + a \rangle$ slip. The immobile dislocations do not contribute to plastic straining, and instead act as ‘forest’ dislocations that impede the motion of dislocations on all the easy-glide/slip systems. Plastic flow on slip system α (where α indicates basal, prism, twin, pyramidal) is controlled by the densities of mobile dislocations ρ_m^α and ‘forest’ dislocations ρ_f^α impeding slip on each slip system. The contribution to the total plastic strain rate from slip system α is then

$$\dot{\epsilon}_p^\alpha = \frac{\rho_m^\alpha b^\alpha}{\sqrt{\sum_{\alpha'} \rho_f^{\alpha'}}} R_{\text{esc}} \quad (1)$$

where

$$R_{\text{esc}} = v_{\text{esc}} \exp \left\{ -\frac{\Delta G_0}{kT} \left[1 - \left(\frac{\tau^\alpha}{\tau_{\text{crit}}^\alpha} \right)^p \right]^q \right\} \quad (2)$$

is the rate of thermally activated escape of the α dislocations past the forest obstacles²⁹. The parameters in equation (2) (attempt frequency v_{esc} , zero-stress activation energy ΔG_0 , and exponents p and q characterizing the activation energy profile) are not important for the general discussion here. The key quantities are the applied resolved shear stress τ^α acting on the dislocations, and the zero-temperature strength or critical flow stress $\tau_{\text{crit}}^\alpha$ required to overcome the energy barrier without thermal activation. $\tau_{\text{crit}}^\alpha$ is due to the forest obstacles, and so is related to the dislocation densities in all slip systems α' including α itself as^{30,31}

$$\tau_{\text{crit}}^\alpha = \mu b^\alpha \sqrt{\sum_{\alpha'} A^{\alpha\alpha'} \rho_f^{\alpha'}} \quad (3)$$

where μ is the shear modulus (ignoring elastic anisotropy) and $A^{\alpha\alpha'}$ is the matrix of interaction strengths between slip systems α and α' . Around room temperature, the fast transition of easy-glide $\langle c + a \rangle$ dislocations into immobile basal-dissociated $\langle c + a \rangle$ or $\langle c \rangle$ dislocations acts to (1) rapidly decrease the density of mobile $\langle c + a \rangle$ dislocations ρ_m^α and (2) rapidly increase the density of immobile/forest dislocations ρ_f^α affecting all slip systems. For all slip systems, the most important effect is the exponential decrease in the escape rate (equation (2)) due to the increased critical strengths on all slip systems (equation (3)) due, in turn, to the increased forest density associated with the transformed $\langle c + a \rangle$ dislocations. The increased forest density also decreases the slip rate directly in equation (1). Therefore, constant-strain-rate experiments show a very rapid increase in the applied stress required to sustain the imposed loading rate, and thus a rapidly increasing hardening rate in the stress-strain curve. High stresses then drive fracture, which cannot be resisted by plastic flow³², and the ductility is thus limited.

Strain hardening is particularly dramatically enhanced in crystals oriented preferentially for $\langle c + a \rangle$ slip. In such an orientation, previously transformed $\langle c + a \rangle$ dislocations on basal planes block subsequent easy-glide pyramidal II $\langle c + a \rangle$ dislocations at their leading edge segments, and drive those dislocations to evolve into long straight segments with pure edge character. These dislocations then transform into immobile dislocations, thus forming immobile dislocation pile-ups. There is thus a feedback process: the rate of transformation increases with increasing immobile dislocation density, while the immobile dislocation density increases owing to the increasing number of transformations. This is observed in single crystal Mg c -axis compression tests^{9,14,19} where the exceptionally high measured work hardening is accompanied by the observation of a rapid increase of straight $\langle c + a \rangle$ dislocations and also loops (density of $\sim 10^{20} \text{ m}^{-3}$ at 1% plastic strain¹⁴). In fact, when $\langle c + a \rangle$ dislocations are observed in TEM^{8,9,11,19,21}, they often exist as uniform arrays or pile-ups of long, straight dislocation segments. In contrast, at low temperatures, the $\langle c + a \rangle$ transitions cannot occur fast enough, leading to more normal

evolution of dislocation densities, strengthening, and strain hardening, although now limited simply by the low temperature. Thus, the pyramidal II $\langle c + a \rangle$ dislocation transformations are responsible for the anomalous temperature dependence of the strengthening observed in the range ~ 130 – 293 K (refs 16, 17), for the high hardening rate observed at 300 K (refs 9, 14, 18, 19), and for the associated low ductility.

Discussion

The transitions identified here are intrinsic to Mg and occur without additional defects or spurious experimental conditions (such as electron beam heating). To prevent the undesirable transitions, strategies could be aimed at energetically stabilizing the easy-glide $\langle c + a \rangle$ core so as to shift the transition to higher temperatures, longer times, or slower strain rates. This may be possible by solute additions that pin the easy-glide $\langle c + a \rangle$ core and lower its energy. Quantitative models demonstrate that glide dislocations are pinned by favourable statistical fluctuations of the solute distribution³³, and the favourable fluctuations for the easy-glide $\langle c + a \rangle$ core will not simultaneously be favourable for the basal-dissociated core at the same position. New encouraging results^{11,34,35} on Mg–3%Y alloys show substantially increased $\langle c + a \rangle$ activity and higher ductility, which may be due to increased stability of the easy-glide $\langle c + a \rangle$ produced by these solutes. Modelling of the interaction between alloying elements and the various dislocation cores found here, and analysis of the consequent effects on the stability of the pyramidal II $\langle c + a \rangle$ edge dislocation, would thus appear to be a useful direction to pursue. Our results also suggest that sufficiently small grain sizes could be favourable for ductility. If easy-glide $\langle c + a \rangle$ dislocations exist long enough to fully traverse grains, then the undesirable transitions are avoided. Indeed, TEM observations in large-grain materials show arrays of straight $\langle c + a \rangle$ dislocation segments^{11,21} inside grains but particularly near grain boundaries³⁶, suggesting that the $\langle c + a \rangle$ are nucleated near grain boundaries but then only travel a short distance before transforming. In contrast, high ductility is achieved in Mg with micrometre-scale grain sizes¹⁰. Finally, the mechanism of the $\langle c + a \rangle$ transition and its consequent strength anomalies may not be unique to Mg; similar observations are seen in other hcp metals^{7,17,20} (such as Cd and Zn). Since the observed transition is intrinsic, results in immobile dislocations, and occurs at high rates, it may also provide the ‘unknown’ pinning process invoked⁷ in one proposed mechanism aimed at rationalizing the observation of basal-oriented $\langle c + a \rangle$ dislocation loops.

In summary, use of a new DFT-validated interatomic potential in long-time MD studies reveals a rich set of intrinsic structural transitions of the key $\langle c + a \rangle$ dislocations in Mg that explain long-standing experimental puzzles and are responsible for low ductility in Mg. The easy-glide pyramidal II $\langle c + a \rangle$ undergoes thermally activated, stress-dependent transitions into various lower-energy products lying on basal planes. The dislocation structures are in good agreement with experimental observations, the differences between experiments are explained, the temperature range where the transition is operative agrees with experiments, and the product dislocations are immobile and so cause high strain hardening by serving as obstacles for all other dislocations, leading to low ductility. This new overall understanding opens opportunities for design of Mg-based alloys based on the mechanistic concept of energetically stabilizing the easy-glide $\langle c + a \rangle$ dislocations on pyramidal II planes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 April; accepted 27 July 2015.

Published online 21 September 2015.

1. Miller, W. S. *et al.* Recent development in aluminium alloys for the automotive industry. *Mater. Sci. Eng. A* **280**, 37–49 (2000).

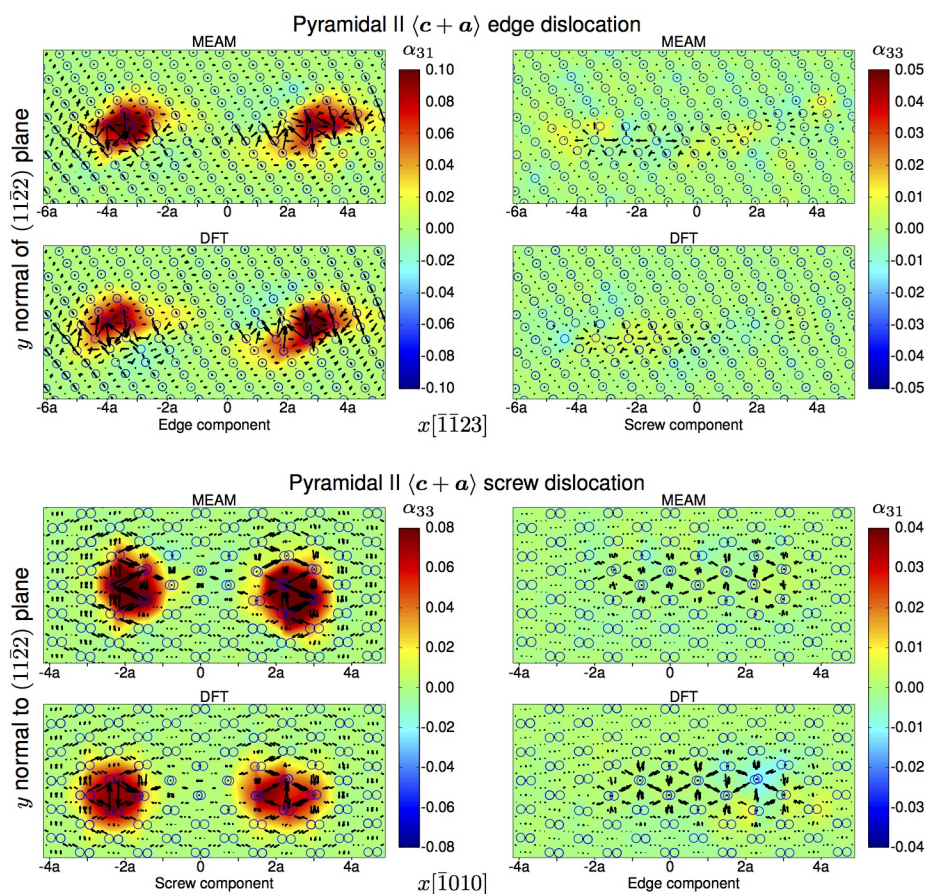
2. Kulecki, M. K. Magnesium and its alloys applications in automotive industry. *Int. J. Adv. Manuf. Technol.* **39**, 851–865 (2008).
3. Heinz, A. *et al.* Recent development in aluminium alloys for aerospace applications. *Mater. Sci. Eng. A* **280**, 102–107 (2000).
4. Pollock, T. M. Weight loss with magnesium alloys. *Science* **328**, 986–987 (2010).
5. Hirth, J. P. & Lothe, J. *Theory of Dislocations* 2nd edn (Wiley & Sons, 1982).
6. Agnew, S. R. *Deformation Mechanisms of Magnesium Alloys* Ch. 2 63–104 (Woodhead Publishing, 2012).
7. Price, P. B. Nonbasal glide in dislocation-free cadmium crystals. II. The $\{11\bar{2}2\}\{11\bar{2}3\}$ system. *J. Appl. Phys.* **32**, 1750–1757 (1961).
8. Stohr, J. F. & Poirier, J. P. Etude en microscopie électronique du glissement pyramidal $\{11\bar{2}2\}\{11\bar{2}3\}$ dans le magnésium. *Phil. Mag.* **25**, 1313–1329 (1972).
9. Obara, T., Yoshinga, H. & Morozumi, S. $\{11\bar{2}2\}\{11\bar{2}3\}$ slip system in magnesium. *Acta Metall.* **21**, 845–853 (1973).
10. Koike, J. *et al.* The activity of non-basal slip systems and dynamic recovery at room temperature in fine-grained AZ31B magnesium alloys. *Acta Mater.* **51**, 2055–2065 (2003).
11. Sandlöbes, S., Friák, M., Neugebauer, J. & Raabe, D. Basal and non-basal dislocation slip in Mg–Y. *Mater. Sci. Eng. A* **576**, 61–68 (2013).
12. Yang, Z., Chisholm, M. F., Duscher, G., Ma, X. & Pennycook, S. J. Direct observation of dislocation dissociation and Suzuki segregation in a Mg–Zn–Y alloy by aberration-corrected scanning transmission electron microscopy. *Acta Mater.* **61**, 350–359 (2013).
13. Yu, Q., Qi, L., Mishra, R. K., Li, J. & Minor, A. M. Reducing deformation anisotropy to achieve ultrahigh strength and ductility in Mg at the nanoscale. *Proc. Natl Acad. Sci. USA* **110**, 13289–13293 (2013).
14. Geng, J., Chisholm, M. F., Mishra, R. K. & Kumar, K. S. The structure of $\langle c + a \rangle$ type dislocation loops in magnesium. *Phil. Mag. Lett.* **94**, 377–386 (2014).
15. Kitahara, T., Ando, S., Tsushida, M., Kitahara, H. & Tonda, H. Deformation behavior of magnesium single crystals in c-axis compression. *Key Eng. Mater.* **345–346**, 129–132 (2007).
16. Ando, S. & Tonda, H. Non-basal slips in magnesium and magnesium-lithium alloy single crystals. *Mater. Sci. Forum* **350–351**, 43–48 (2000).
17. Tonda, H. & Ando, S. Effect of temperature and shear direction on yield stress by $\{11\bar{2}2\}\{11\bar{2}3\}$ slip in HCP metals. *Metall. Mater. Trans. A* **33**, 831–836 (2002).
18. Byer, C. M., Li, B., Cao, B. & Ramesh, K. T. Microcompression of single-crystal magnesium. *Scr. Mater.* **62**, 536–539 (2010).
19. Syed, B., Geng, J., Mishra, R. K. & Kumar, K. S. $[0001]$ compression response at room temperature of single-crystal magnesium. *Scr. Mater.* **67**, 700–703 (2012).
20. Price, P. B. Pyramidal glide and the formation and climb of dislocation loops in nearly perfect zinc crystals. *Phil. Mag.* **5**, 873–886 (1960).
21. Agnew, S. R., Horton, J. A. & Yoo, M. H. Transmission electron microscopy investigation of $\langle c + a \rangle$ dislocations in Mg and α -solid solution Mg–Li alloys. *Metall. Mater. Trans. A* **33**, 851–858 (2002).
22. Bettles, C. J., Gibson, M. A. & Zhu, S. M. Microstructure and mechanical behaviour of an elevated temperature Mg–rare earth based alloy. *Mater. Sci. Eng. A* **505**, 6–12 (2009).
23. Hidalgo-Manrique, P. *et al.* Origin of the reversed yield asymmetry in Mg–rare earth alloys at high temperature. *Acta Mater.* **92**, 265–277 (2015).
24. Yoo, M. H., Agnew, S. R., Morris, J. R. & Ho, K. M. Non-basal slip systems in HCP metals and alloys: source mechanisms. *Mater. Sci. Eng. A* **319–321**, 87–92 (2001).
25. Edelin, G. & Poirier, J. P. Etude de la montée des dislocations au moyen d'expériences de fluage par diffusion dans le magnésium. *Phil. Mag.* **28**, 1203–1210 (1973).
26. Wu, Z., Francis, M. F. & Curtin, W. A. Magnesium interatomic potential for simulating plasticity and fracture phenomena. *Model. Simul. Mater. Sci. Eng.* **23**, 015004 (2015).
27. Henkelman, G. & Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **113**, 9978–9985 (2000).
28. Henkelman, G., Uberuaga, B. P. & Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **113**, 9901–9904 (2000).
29. Kocks, U. F., Argon, A. S. & Ashby, F. Thermodynamics and kinetics of slip. *Prog. Mater. Sci.* **19**, 1–291 (1975).
30. Devincere, B., Hoc, T. & Kubin, L. Dislocation mean free paths and strain hardening of crystals. *Science* **320**, 1745–1748 (2008).
31. Bertin, N., Tomé, C. N., Beyerlein, I. J., Barnett, M. R. & Capolungo, L. On the strength of dislocation interactions and their effect on latent hardening in pure magnesium. *Int. J. Plast.* **62**, 72–92 (2014).
32. Wu, Z. & Curtin, W. A. Brittle and ductile crack-tip behavior in magnesium. *Acta Mater.* **88**, 1–12 (2015).
33. Leyson, G. P. M., Curtin, W. A., Hector, L. G. & Woodward, C. F. Quantitative prediction of solute strengthening in aluminium alloys. *Nature Mater.* **9**, 750–755 (2010).
34. Sandlöbes, S. *et al.* The relation between ductility and stacking fault energies in Mg and Mg–Y alloys. *Acta Mater.* **60**, 3011–3021 (2012).
35. Sandlöbes, S. *et al.* Ductility improvement of Mg alloys by solid solution: ab initio modeling, synthesis and mechanical properties. *Acta Mater.* **70**, 92–104 (2014).
36. Kang, F., Liu, J. Q., Wang, J. T. & Zhao, X. The effect of hydrostatic pressure on the activation of non-basal slip in a magnesium alloy. *Scr. Mater.* **61**, 844–847 (2009).
37. Faken, D. & Jónsson, H. Systematic analysis of local atomic structure combined with 3D computer graphics. *Comput. Mater. Sci.* **2**, 279–286 (1994).

Supplementary Information is available in the online version of the paper.

Acknowledgements Z.W. acknowledges financial support from the Agency for Science, Technology and Research (A*STAR), Singapore. W.A.C. acknowledges support of this work through a European Research Council Advanced Grant, 'Predictive Computational Metallurgy', ERC grant agreement no. 339081 – PreCoMet. W.A.C. also acknowledges earlier long-term support of Mg research from General Motors Corporation that provided the basis for research reported here.

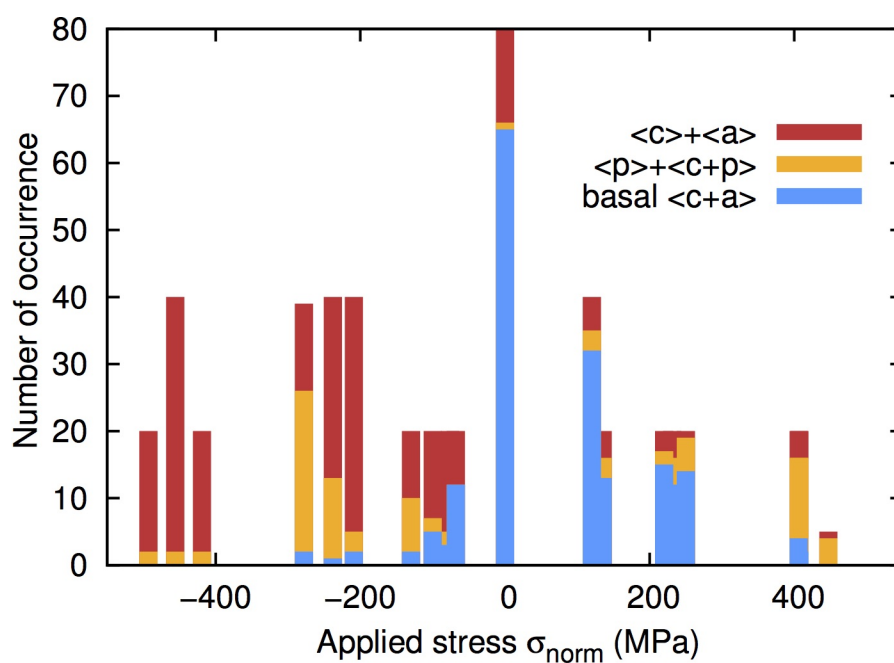
Author Contributions Z.W. and W.A.C. designed the research, analysed the data, developed the model, discussed the results, and wrote the paper. Z.W. performed the molecular dynamics simulations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.A.C. (william.curtin@epfl.ch).

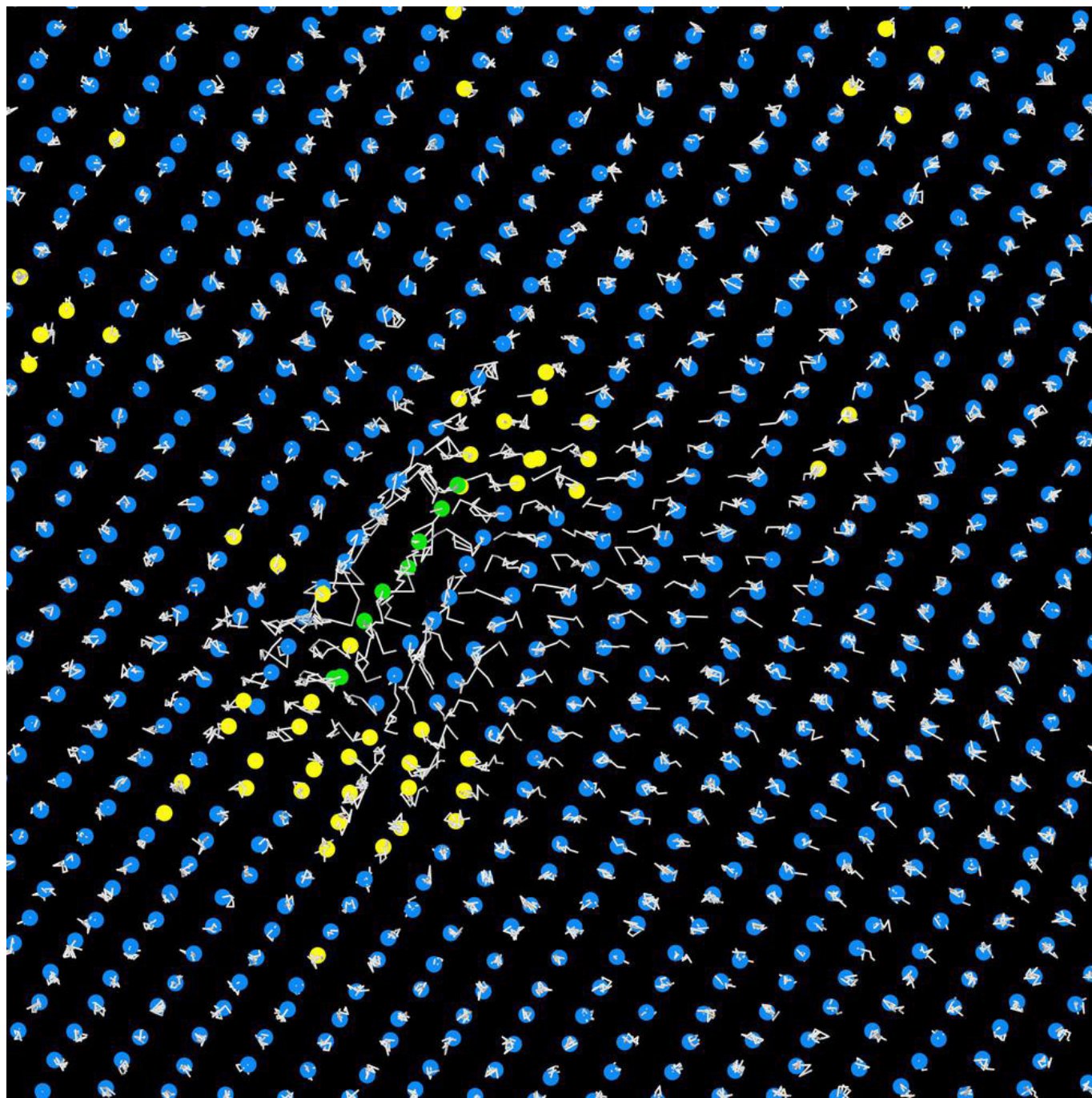


Extended Data Figure 1 | Dislocation core structures. Pyramidal II $\langle c + a \rangle$ edge dislocation (top) and screw dislocation (bottom) core structures predicted by the modified embedded-atom method (MEAM) potential and DFT as visualized by the component of the Nye tensor and differential displacement plots²⁶. The circles depict atoms projected onto the plane perpendicular to the dislocation line direction with the crystallographic orientation of the x and y axes shown (a is the lattice parameter of the hcp unit cell). In each image, the distribution of colour represents the distribution of

the Nye tensor component (in units of \AA^{-1}), that is, the distribution of infinitesimal dislocation Burgers vector; the arrows represent the relative displacement component between two neighbouring atoms. At the dislocation cores, similarities in the distributions of colour and patterns of arrows between MEAM and DFT suggest similarities in atomic structures predicted by the two models. Figure from ref. 26 (<http://dx.doi.org/10.1088/0965-0393/23/1/015004>), copyright IOP Publishing. Reproduced with permission. All rights reserved.

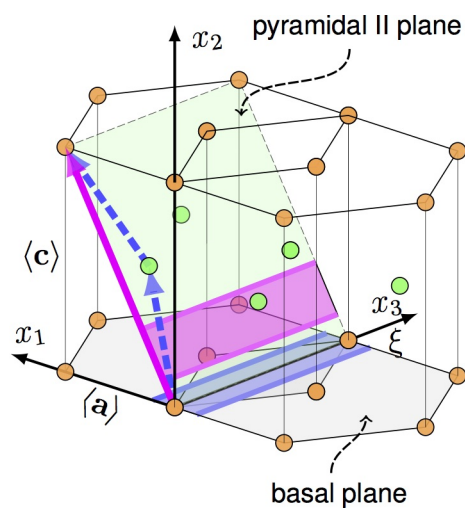


Extended Data Figure 2 | Distribution of pyramidal $\langle c+a \rangle$ edge dislocation transition type at different applied stresses. At zero or low applied stresses, basal $\langle c+a \rangle$ is dominant (blue), while partial $\langle a \rangle$ and $\langle c \rangle$ (orange) or full $\langle a \rangle$ and $\langle c \rangle$ (red) dislocations are dominant at high applied stresses.

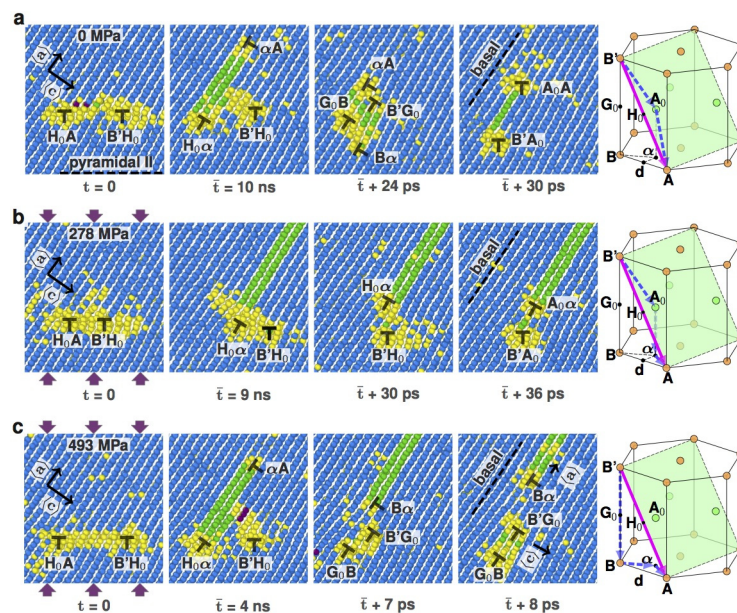


Extended Data Figure 3 | Atom trajectories during transition from pyramidal II $\langle c + a \rangle$ to basal $\langle c + a \rangle$ at 500 K. Each white line traces individual atom trajectory during the transition. No vacancy/interstitial diffusion from the

bulk is involved during the transition. Atoms and trajectories are projected onto the plane perpendicular to the dislocation line direction (see Fig. 1 caption for colour representation).

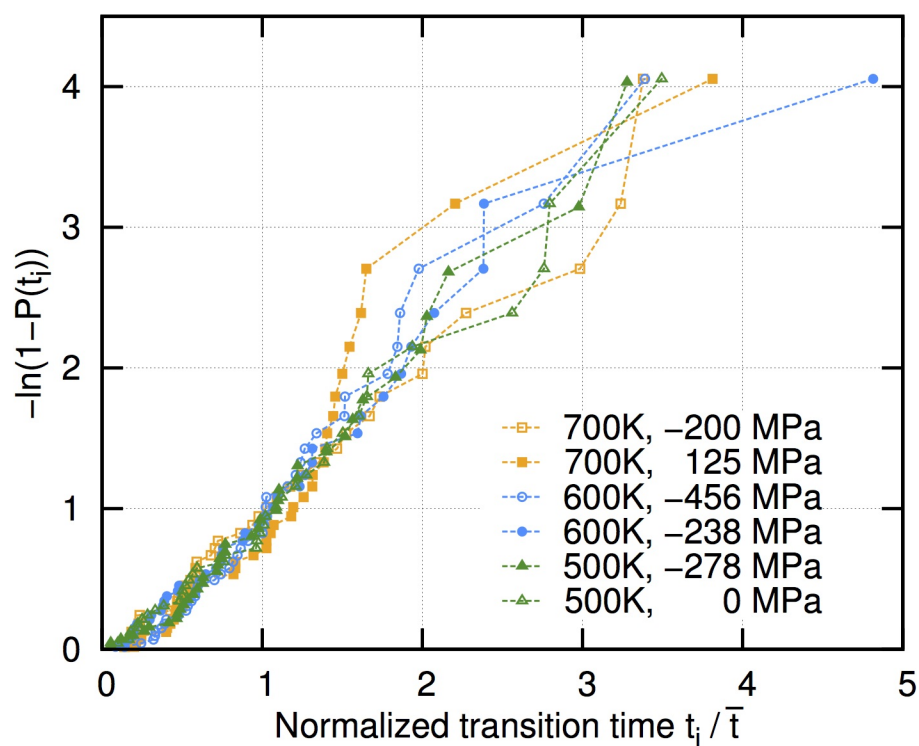


Extended Data Figure 4 | Schematics and coordinate system for $\langle c+a \rangle$ dislocation transition. The $\langle c+a \rangle$ edge dislocation dissociated with a stacking fault (shown pink) on the pyramidal II plane (shown green) climb-dissociates into the basal plane with an I_1 stacking fault (shown blue). The pink solid arrow and blue dashed arrows indicate the relevant $\langle c+a \rangle$ Burgers vectors before and after climb-dissociation. ξ indicates the dislocation line direction and x_1 , x_2 , and x_3 are the Cartesian coordinate system used for calculating dislocation elastic energy.

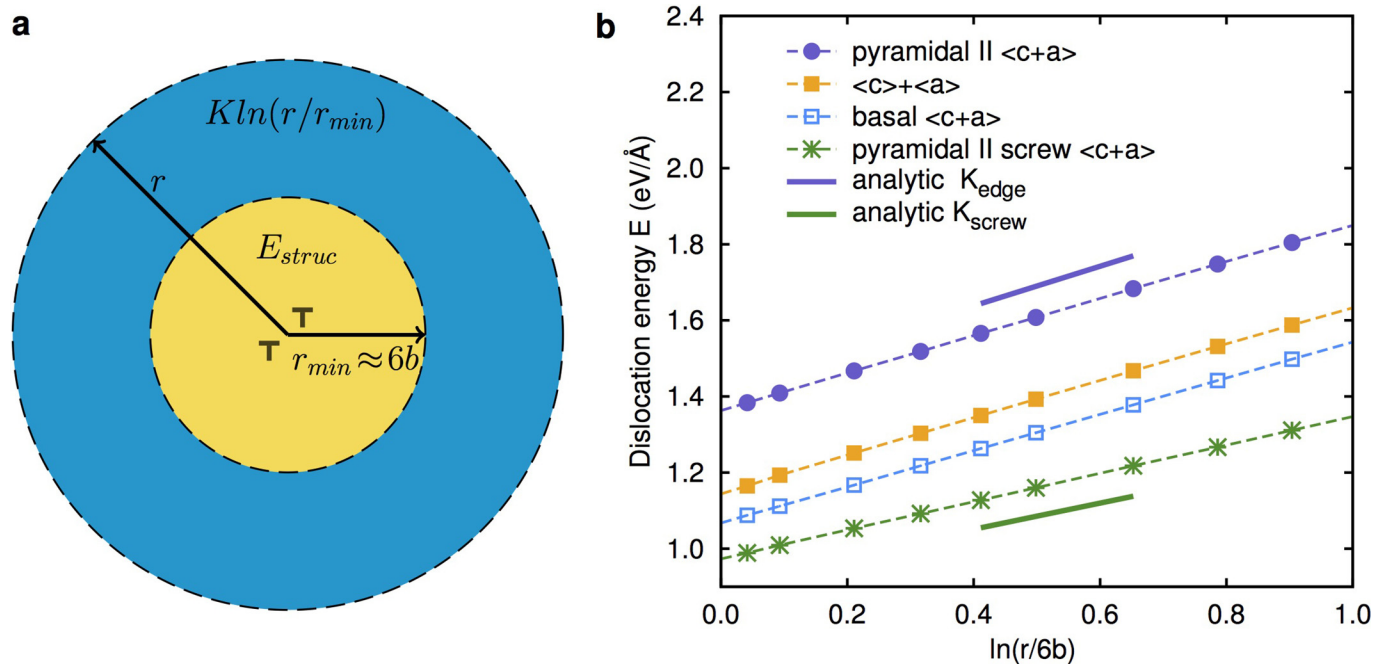


Extended Data Figure 5 | Pyramidal II $\langle c + a \rangle$ edge dislocation dissociation. Shown are details of the dissociation of the easy-glide pyramidal II $\langle c + a \rangle$ into basal-dissociated products as observed during long-time MD simulations, for **a**, zero, **b**, moderate, and **c**, high compressive stresses normal to

the pyramidal II plane. Dislocation cores are indicated by the symbol '⊥'. See Supplementary Information V for details of the dislocation reactions and Burgers vectors.

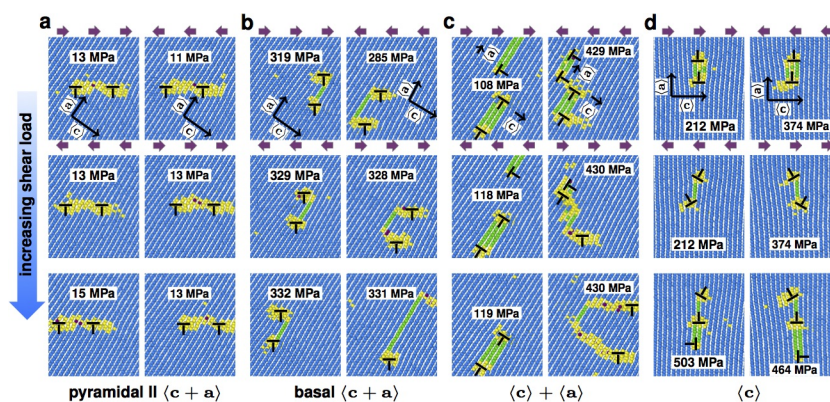


Extended Data Figure 6 | Transition probability distribution for the random $\langle c + a \rangle$ transition processes. P is the cumulative transition probability, \bar{t} is the mean transition time, and t_i indicates the ordered transition times from smallest to largest, $t_1 < t_2 < \dots < t_N$. See Supplementary Information VII for details.



Extended Data Figure 7 | Energy of $\langle c+a \rangle$ dislocations calculated within a cylindrical region of radius r . **a**, Schematic showing the total dislocation energy consisting of near-core energy E_{struc} and far-field elastic energy $K \ln(r/r_{min})$, where $r_{min} \approx 6b = 6|\langle c+a \rangle|$. **b**, The four dislocation energies (dashed lines, top to bottom) correspond respectively to the $\langle c+a \rangle$ edge dislocation on the pyramidal II plane, edge $\langle c \rangle$ and $\langle a \rangle$ in close proximity, $\langle c+a \rangle$ edge dislocation climb-dissociated on the basal plane, and the

$\langle c+a \rangle$ screw dislocation on the pyramidal II plane. The edge dislocations with different core configurations have different energy density within r_{min} but the same far-field elastic energy scaling K . The screw dislocation has both lower energy within r_{min} and lower elastic energy scaling than do all the edge dislocations. The analytical energy prefactors K (the slope) from the anisotropic elastic solution are also shown (solid lines).



Extended Data Figure 8 | Glide behaviour of the various dislocations under resolved shear stresses (directions indicated) at 300 K. **a**, Easy-glide pyramidal II $\langle c + a \rangle$ at ~ 13 MPa and 11 MPa with a distinct directional dependence. **b**, Glide of the basal-dissociated $\langle c + a \rangle$ but at very high stresses,

~ 330 MPa. **c**, Glide of the $\langle a \rangle$ dislocation away from the remaining $\langle c + a \rangle$ product at ~ 119 MPa, leaving an immobile $\langle c \rangle$, and reaction of $\langle a \rangle$ and $\langle c \rangle$ forming the basal $\langle c + a \rangle$ dislocation. **d**, Nucleation of partial $\langle a \rangle$ dislocations at 400–600 MPa from the immobile basal-dissociated $\langle c \rangle$ dislocation.

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

The 1000 Genomes Project has already elucidated the properties and distribution of common and rare variation, provided insights into the processes that shape genetic diversity, and advanced understanding of disease biology^{1,2}. This resource provides a benchmark for surveys of human genetic variation and constitutes a key component for human genetic studies, by enabling array design^{3,4}, genotype imputation⁵, cataloguing of variants in regions of interest, and filtering of likely neutral variants^{6,7}.

In this final phase, individuals were sampled from 26 populations in Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and the Americas (AMR) (Fig. 1a; see Supplementary Table 1 for population descriptions and abbreviations). All individuals were sequenced using both whole-genome sequencing (mean depth = 7.4×) and targeted exome sequencing (mean depth = 65.7×). In addition, individuals and available first-degree relatives (generally, adult offspring) were genotyped using high-density SNP microarrays. This provided a cost-effective means to discover genetic variants and estimate individual genotypes and haplotypes^{1,2}.

Data set overview

In contrast to earlier phases of the project, we expanded analysis beyond bi-allelic events to include multi-allelic SNPs, indels, and a diverse set of structural variants (SVs). An overview of the sample collection, data generation, data processing, and analysis is given in Extended Data Fig. 1. Variant discovery used an ensemble of 24 sequence analysis tools (Supplementary Table 2), and machine-learning classifiers to separate high-quality variants from potential false positives, balancing sensitivity and specificity. Construction of haplotypes started with estimation of long-range phased haplotypes using array genotypes for project participants and, where available, their first degree relatives; continued with the addition of high confidence bi-allelic variants that were analysed jointly to improve these haplotypes; and concluded with the placement of multi-allelic and structural variants onto the haplotype scaffold one at a time (Box 1). Overall, we discovered, genotyped, and phased 88 million variant sites (Supplementary Table 3). The project has now contributed or validated 80 million of the 100 million variants in the public dbSNP catalogue (version 141 includes 40 million SNPs and indels newly

contributed by this analysis). These novel variants especially enhance our catalogue of genetic variation within South Asian (which account for 24% of novel variants) and African populations (28% of novel variants).

To control the false discovery rate (FDR) of SNPs and indels at <5%, a variant quality score threshold was defined using high depth (>30×) PCR-free sequence data generated for one individual per population. For structural variants, additional orthogonal methods were used for confirmation, including microarrays and long-read sequencing, resulting in FDR < 5% for deletions, duplications, multi-allelic copy-number variants, Alu and L1 insertions, and <20% for inversions, SVA (SINE/VNTR/Alu) composite retrotransposon insertions and NUMTs⁸ (nuclear mitochondrial DNA variants). To evaluate variant discovery power and genotyping accuracy, we also generated deep Complete Genomics data (mean depth = 47×) for 427 individuals (129 mother–father–child trios, 12 parent–child duos, and 16 unrelateds). We estimate the power to detect SNPs and indels to be >95% and >80%, respectively, for variants with sample frequency of at least 0.5%, rising to >99% and >85% for frequencies >1% (Extended Data Fig. 2). At lower frequencies, comparison with >60,000 European haplotypes from the Haplotype Reference Consortium⁹ suggests 75% power to detect SNPs with frequency of 0.1%. Furthermore, we estimate heterozygous genotype accuracy at 99.4% for SNPs and 99.0% for indels (Supplementary Table 4), a threefold reduction in error rates compared to our previous release², resulting from the larger sample size, improvements in sequence data accuracy, and genotype calling and phasing algorithms.

A typical genome

We find that a typical genome differs from the reference human genome at 4.1 million to 5.0 million sites (Fig. 1b and Table 1). Although >99.9% of variants consist of SNPs and short indels, structural variants affect more bases: the typical genome contains an estimated 2,100 to 2,500 structural variants (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), affecting ~20 million bases of sequence.

*Lists of participants and their affiliations appear in the online version of the paper.

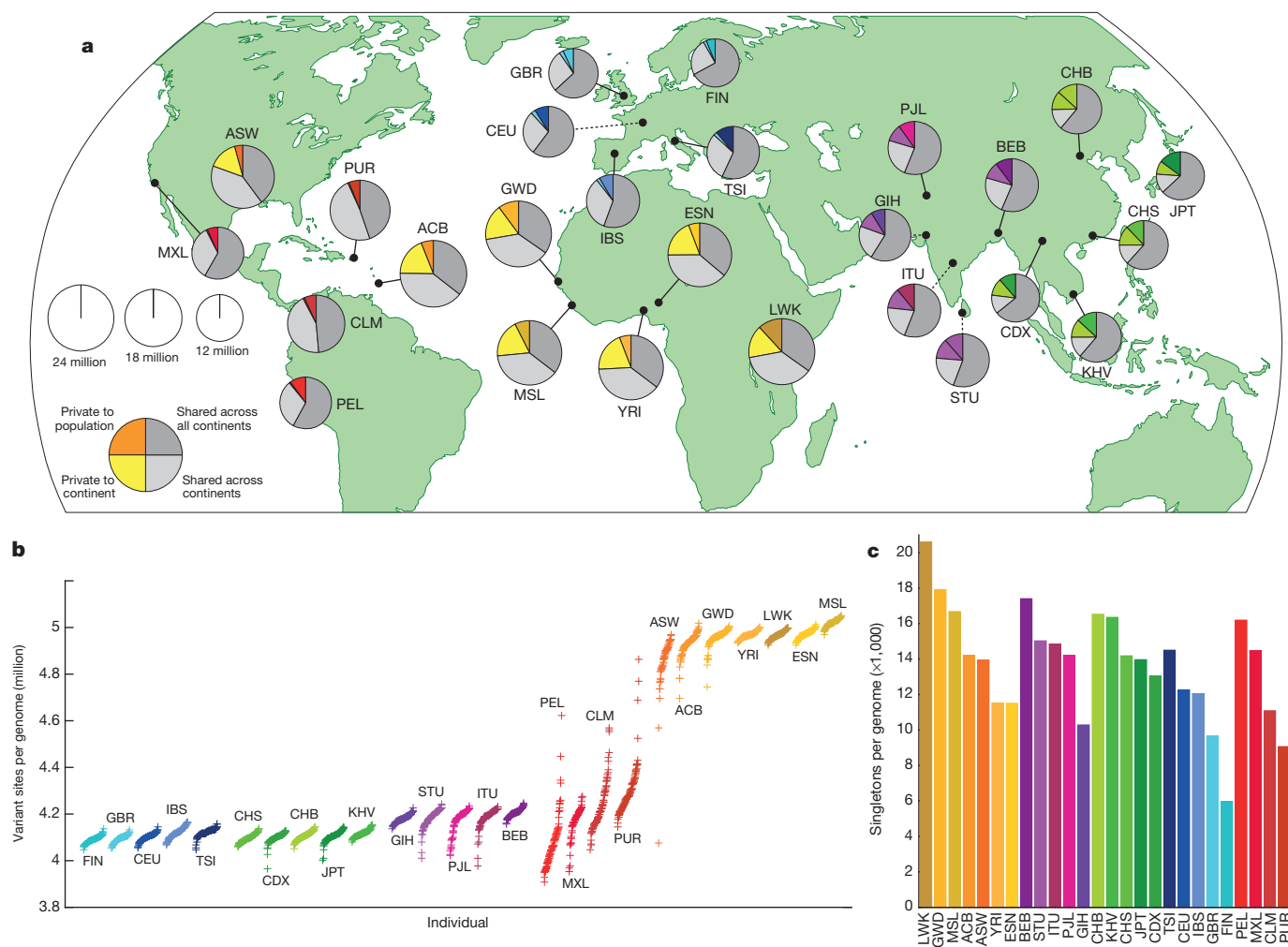


Figure 1 | Population sampling. **a**, Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private to a continental area (lighter colour shared across continental group), shared

across continental areas (light grey), and shared across all continents (dark grey). Dashed lines indicate populations sampled outside of their ancestral continental region. **b**, The number of variant sites per genome. **c**, The average number of singletons per genome.

The total number of observed non-reference sites differs greatly among populations (Fig. 1b). Individuals from African ancestry populations harbour the greatest numbers of variant sites, as predicted by the out-of-Africa model of human origins. Individuals from recently admixed populations show great variability in the number of variants, roughly proportional to the degree of recent African ancestry in their genomes.

The majority of variants in the data set are rare: ~64 million autosomal variants have a frequency <0.5%, ~12 million have a frequency between 0.5% and 5%, and only ~8 million have a frequency >5% (Extended Data Fig. 3a). Nevertheless, the majority of variants observed in a single genome are common: just 40,000 to 200,000 of the variants in a typical genome (1–4%) have a frequency <0.5% (Fig. 1c and Extended Data Fig. 3b). As such, we estimate that improved rare variant discovery by deep sequencing our entire sample would at least double the total number of variants in our sample but increase the number of variants in a typical genome by only ~20,000 to 60,000.

Putatively functional variation

When we restricted analyses to the variants most likely to affect gene function, we found a typical genome contained 149–182 sites with protein truncating variants, 10,000 to 12,000 sites with peptide-sequence-altering variants, and 459,000 to 565,000 variant sites overlapping known regulatory regions (untranslated regions (UTRs),

promoters, insulators, enhancers, and transcription factor binding sites). African genomes were consistently at the high end of these ranges. The number of alleles associated with a disease or phenotype in each genome did not follow this pattern of increased diversity in Africa (Extended Data Fig. 4): we observed ~2,000 variants per genome associated with complex traits through genome-wide association studies (GWAS) and 24–30 variants per genome implicated in rare disease through ClinVar; with European ancestry genomes at the high-end of these counts. The magnitude of this difference is unlikely to be explained by demography^{10,11}, but instead reflects the ethnic bias of current genetic studies. We expect that improved characterization of the clinical and phenotypic consequences of non-European alleles will enable better interpretation of genomes from all individuals and populations.

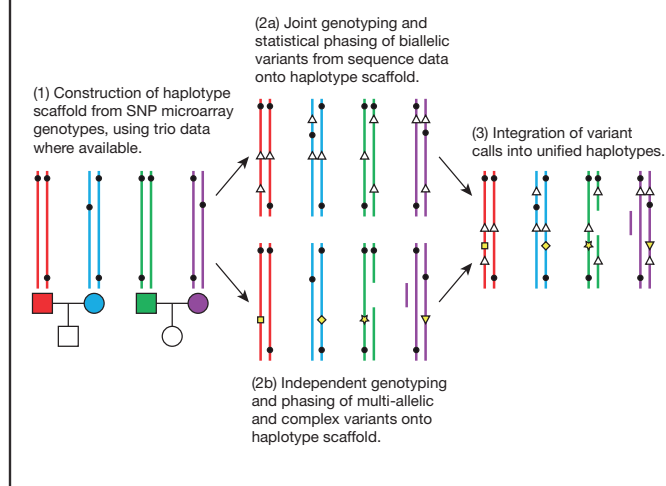
Sharing of genetic variants among populations

Systematic analysis of the patterns in which genetic variants are shared among individuals and populations provides detailed accounts of population history. Although most common variants are shared across the world, rarer variants are typically restricted to closely related populations (Fig. 1a); 86% of variants were restricted to a single continental group. Using a maximum likelihood approach¹², we estimated the proportion of each genome derived from several putative ‘ancestral populations’ (Fig. 2a and Extended Data Fig. 5).

BOX 1

Building a haplotype scaffold

To construct high quality haplotypes that integrate multiple variant types, we adopted a staged approach³⁷. (1) A high-quality 'haplotype scaffold' was constructed using statistical methods applied to SNP microarray genotypes (black circles) and, where available, genotypes for first degree relatives (available for ~52% of samples; Supplementary Table 11)³⁸. (2a) Variant sites were identified using a combination of bioinformatic tools and pipelines to define a set of high-confidence bi-allelic variants, including both SNPs and indels (white triangles), which were jointly imputed onto the haplotype scaffold. (2b) Multi-allelic SNPs, indels, and complex variants (represented by yellow shapes, or variation in copy number) were placed onto the haplotype scaffold one at a time, exploiting the local linkage disequilibrium information but leaving haplotypes for other variants undisturbed³⁹. (3) The biallelic and multi-allelic haplotypes were merged into a single haplotype representation. This multi-stage approach allows the long-range structure of the haplotype scaffold to be maintained while including more complex types of variation. Comparison to haplotypes constructed from fosmids suggests the average distance between phasing errors is ~1,062 kb, with typical phasing errors stretching ~37 kb (Supplementary Table 12).



This analysis separates continental groups, highlights their internal substructure, and reveals genetic similarities between related populations. For example, east–west clines are visible in Africa and East Asia, a north–south cline is visible in Europe, and European, African, and Native-American admixture is visible in genomes sampled in the Americas.

To characterize more recent patterns of shared ancestry, we first focused on variants observed on just two chromosomes (sample frequency of 0.04%), the rarest shared variants within our sample, and known as f_2 variants². As expected, these variants are typically geographically restricted and much more likely to be shared between individuals in the same population or continental group, or between populations with known recent admixture (Extended Data Fig. 6a, b). Analysis of shared haplotype lengths around f_2 variants suggests a median common ancestor ~296 generations ago (7,410 to 8,892 years ago; Extended Data Fig. 6c, d), although those confined within a population tend to be younger, with a shared common ancestor ~143 generations ago (3,570 to 4,284 years ago)¹³.

Insights about demography

Modelling the distribution of variation within and between genomes can provide insights about the history and demography of our

ancestor populations¹⁴. We used the pairwise sequentially Markovian coalescent (PSMC)¹⁴ method to characterize the effective population size (N_e) of the ancestral populations (Fig. 2b and Extended Data Fig. 7). Our results show a shared demographic history for all humans beyond ~150,000 to 200,000 years ago. Further, they show that European, Asian and American populations shared strong and sustained bottlenecks, all with $N_e < 1,500$, between 15,000 to 20,000 years ago. In contrast, the bottleneck experienced by African populations during the same time period appears less severe, with $N_e > 4,250$. These bottlenecks were followed by extremely rapid inferred population growth in non-African populations, with notable exceptions including the PEL, MXL and FIN.

Due to the shared ancestry of all humans, only a modest number of variants show large frequency differences among populations. We observed 762,000 variants that are rare (defined as having frequency <0.5%) within the global sample but much more common (>5% frequency) in at least one population (Fig. 3a). Several populations have relatively large numbers of these variants, and these are typically genetically or geographically distinct within their continental group (LWK in Africa, PEL in the Americas, JPT in East Asia, FIN in Europe, and GIH in South Asia; see Supplementary Table 5). Drifted variants within such populations may reveal phenotypic associations that would be hard to identify in much larger global samples¹⁵.

Analysis of the small set of variants with large frequency differences between closely related populations can identify targets of recent, localized adaptation. We used the F_{ST} -based population branch statistic (PBS)¹⁶ to identify genes with strong differentiation between pairs of populations in the same continental group (Fig. 3b). This approach reveals a number of previously identified selection signals (such as *SLC24A5* associated with skin pigmentation¹⁷, *HERC2* associated with eye colour¹⁸, *LCT* associated with lactose tolerance, and the *FADS* cluster that may be associated with dietary fat sources¹⁹). Several potentially novel selection signals are also highlighted (such as *TRBV9*, which appears particularly differentiated in South Asia, *PRICKLE4*, differentiated in African and South Asian populations, and a number of genes in the immunoglobulin cluster, differentiated in East Asian populations; Extended Data Fig. 8), although at least some of these signals may result from somatic rearrangements (for example, via V(D)J recombination) and differences in cell type composition among the sequenced samples. Nonetheless, the relatively small number of genes showing strong differentiation between closely related populations highlights the rarity of strong selective sweeps in recent human evolution²⁰.

Sharing of haplotypes and imputation

The sharing of haplotypes among individuals is widely used for imputation in GWAS, a primary use of 1000 Genomes data. To assess imputation based on the phase 3 data set, we used Complete Genomics data for 9 or 10 individuals from each of 6 populations (CEU, CHS, LWK, PEL, PJL, and YRI). After excluding these individuals from the reference panel, we imputed genotypes across the genome using sites on a typical one million SNP microarray. The squared correlation between imputed and experimental genotypes was >95% for common variants in each population, decreasing gradually with minor allele frequency (Fig. 4a). Compared to phase 1, rare variation imputation improved considerably, particularly for newly sampled populations (for example, PEL and PJL, Extended Data Fig. 9a). Improvements in imputations restricted to overlapping samples suggest approximately equal contributions from greater genotype and sequence quality and from increased sample size (Fig. 4a, inset). Imputation accuracy is now similar for bi-allelic SNPs, bi-allelic indels, multi-allelic SNPs, and sites where indels and SNPs overlap, but slightly reduced for multi-allelic indels, which typically map to regions of low-complexity sequence and are much harder to genotype and phase (Extended Data Fig. 9b). Although imputation of rare variation remains challenging, it appears to be

Table 1 | Median autosomal variant sites per genome

| | AFR | | AMR | | EAS | | EUR | | SAS | |
|-----------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Samples | 661 | | 347 | | 504 | | 503 | | 489 | |
| Mean coverage | 8.2 | | 7.6 | | 7.7 | | 7.4 | | 8.0 | |
| | Var. sites | Singletons | Var. sites | Singletons | Var. sites | Singletons | Var. sites | Singletons | Var. sites | Singletons |
| SNPs | 4.31M | 14.5k | 3.64M | 12.0k | 3.55M | 14.8k | 3.53M | 11.4k | 3.60M | 14.4k |
| Indels | 625k | - | 557k | - | 546k | - | 546k | - | 556k | - |
| Large deletions | 1.1k | 5 | 949 | 5 | 940 | 7 | 939 | 5 | 947 | 5 |
| CNVs | 170 | 1 | 153 | 1 | 158 | 1 | 157 | 1 | 165 | 1 |
| MEI (Alu) | 1.03k | 0 | 845 | 0 | 899 | 1 | 919 | 0 | 889 | 0 |
| MEI (L1) | 138 | 0 | 118 | 0 | 130 | 0 | 123 | 0 | 123 | 0 |
| MEI (SVA) | 52 | 0 | 44 | 0 | 56 | 0 | 53 | 0 | 44 | 0 |
| MEI (MT) | 5 | 0 | 5 | 0 | 4 | 0 | 4 | 0 | 4 | 0 |
| Inversions | 12 | 0 | 9 | 0 | 10 | 0 | 9 | 0 | 11 | 0 |
| Nonsynon | 12.2k | 139 | 10.4k | 121 | 10.2k | 144 | 10.2k | 116 | 10.3k | 144 |
| Synon | 13.8k | 78 | 11.4k | 67 | 11.2k | 79 | 11.2k | 59 | 11.4k | 78 |
| Intron | 2.06M | 7.33k | 1.72M | 6.12k | 1.68M | 7.39k | 1.68M | 5.68k | 1.72M | 7.20k |
| UTR | 37.2k | 168 | 30.8k | 136 | 30.0k | 169 | 30.0k | 129 | 30.7k | 168 |
| Promoter | 102k | 430 | 84.3k | 332 | 81.6k | 425 | 82.2k | 336 | 84.0k | 430 |
| Insulator | 70.9k | 248 | 59.0k | 199 | 57.7k | 252 | 57.7k | 189 | 59.1k | 243 |
| Enhancer | 354k | 1.32k | 295k | 1.05k | 289k | 1.34k | 288k | 1.02k | 295k | 1.31k |
| TFBSs | 927 | 4 | 759 | 3 | 748 | 4 | 749 | 3 | 765 | 3 |
| Filtered LoF | 182 | 4 | 152 | 3 | 153 | 4 | 149 | 3 | 151 | 3 |
| HGMD-DM | 20 | 0 | 18 | 0 | 16 | 1 | 18 | 2 | 16 | 0 |
| GWAS | 2.00k | 0 | 2.07k | 0 | 1.99k | 0 | 2.08k | 0 | 2.06k | 0 |
| ClinVar | 28 | 0 | 30 | 1 | 24 | 0 | 29 | 1 | 27 | 1 |

See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.

most accurate in African ancestry populations, where greater genetic diversity results in a larger number of haplotypes and improves the chances that a rare variant is tagged by a characteristic haplotype.

Resolution of genetic association studies

To evaluate the impact of our new reference panel on GWAS, we re-analysed a previous study of age-related macular degeneration (AMD) totalling 2,157 cases and 1,150 controls²¹. We imputed 17.0 million genetic variants with estimated $R^2 > 0.3$, compared to 14.1 million variants using phase 1, and only 2.4 million SNPs using HapMap2. Compared to phase 1, the number of imputed common and intermediate frequency variants increased by 7%, whereas the number of rare variants increased by $>50\%$, and the number of indels increased by 70% (Supplementary Table 6). We permuted case-control labels to estimate a genome-wide significance threshold of $P < \sim 1.5 \times 10^{-8}$, which corresponds to ~ 3 million independent variants and is more stringent than the traditional threshold of 5×10^{-8} (Supplementary Table 7). In practice, significance thresholds must balance false positives and false negatives^{22–24}. We recommend that thresholds aiming for strict control of false positives should be determined using permutations. We expect thresholds to become more stringent when larger sample sizes are sequenced, when diverse samples are studied, or when genotyping and imputation is replaced with direct sequencing. After imputation, five independent signals in four previously reported AMD loci^{25–28} reached genome-wide significance (Supplementary Table 8). When we examined each of these to define a set of potentially causal variants using a Bayesian Credible set approach²⁹, lists of potentially functional variants were $\sim 4\times$ larger than in HapMap2-based analysis and 7% larger than in analyses based on phase 1 (Supplementary Table 9). In the *ARMS2/HTRA1* locus, the most strongly associated variant was now a structural variant (estimated imputation $R^2 = 0.89$) that previously could not be imputed, consistent with some functional studies³⁰. Deep catalogues of potentially functional variants will help ensure that downstream functional analyses include the true candidate variants, and will aid analyses that integrate complex disease associations with functional genomic elements³¹.

The performance of imputation and GWAS studies depends on the local distribution of linkage disequilibrium (LD) between nearby var-

iants. Controlling for sample size, the decay of LD as a function of physical distance is fastest in African populations and slowest in East Asian populations (Extended Data Fig. 10). To evaluate how these differences influence the resolution of genetic association studies and,

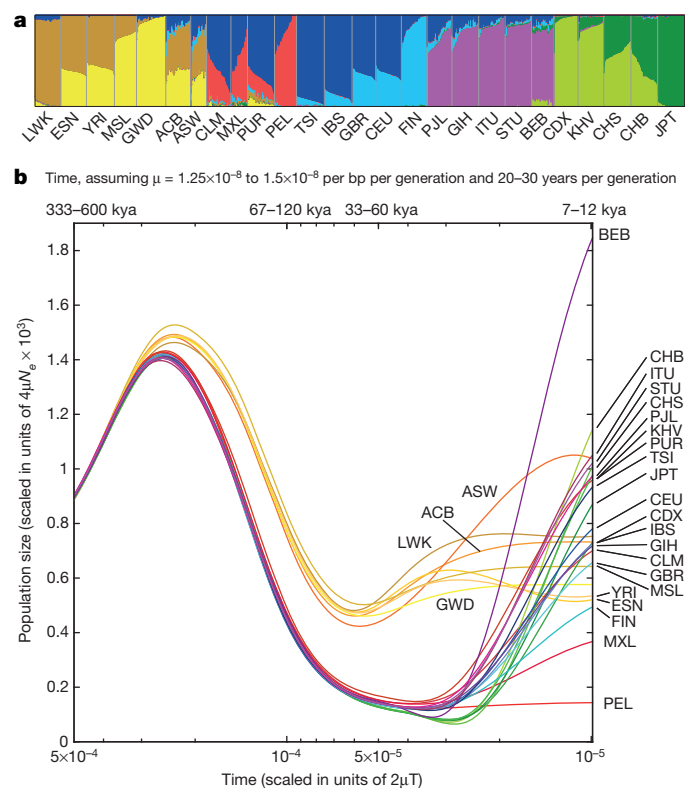


Figure 2 | Population structure and demography. **a**, Population structure inferred using a maximum likelihood approach with 8 clusters. **b**, Changes to effective population sizes over time, inferred using PSMC. Lines represent the within-population median PSMC estimate, smoothed by fitting a cubic spline passing through bin midpoints.

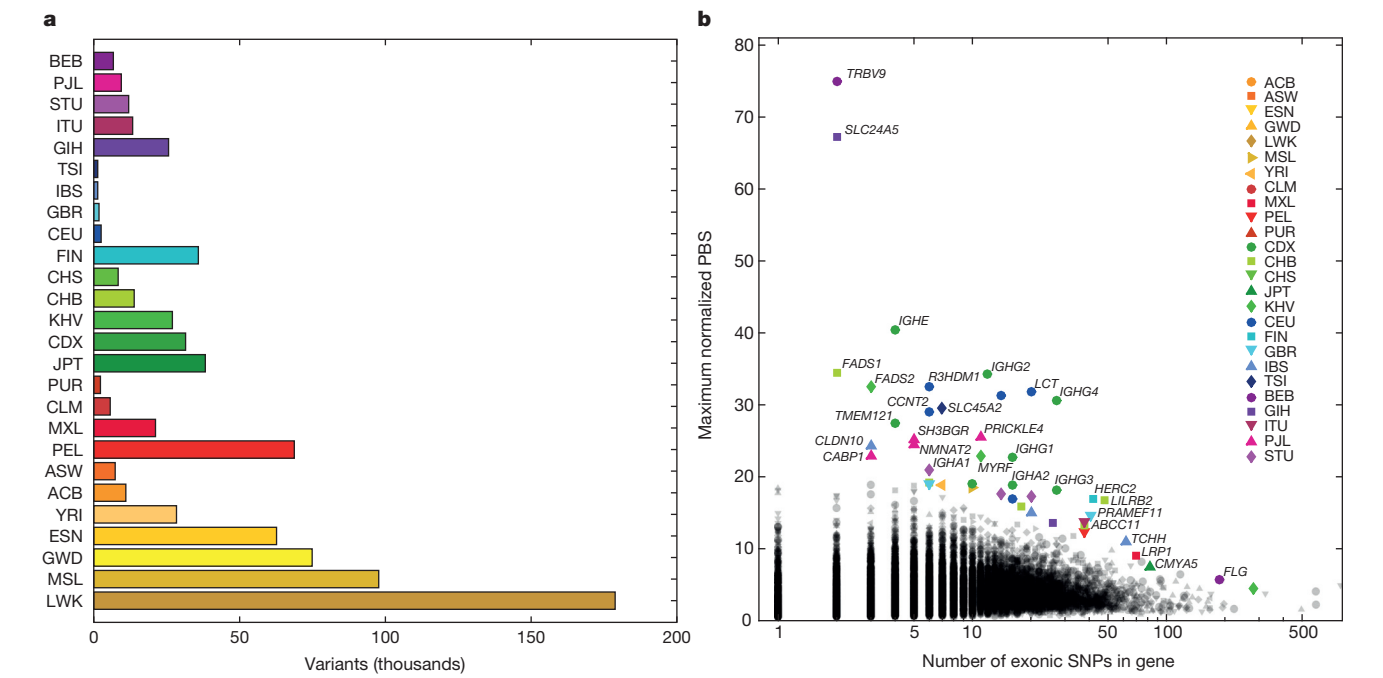


Figure 3 | Population differentiation. **a**, Variants found to be rare ($<0.5\%$) within the global sample, but common ($>5\%$) within a population. **b**, Genes showing strong differentiation between pairs of closely related populations.

in particular, their ability to identify a narrow set of candidate functional variants, we evaluated the number of tagging variants ($r^2 > 0.8$) for a typical variant in each population. We find that each common variant typically has over 15–20 tagging variants in non-African populations, but only about 8 in African populations (Fig. 4b). At lower frequencies, we find 3–6 tagging variants with 100 kb of variants

The vertical axis gives the maximum obtained value of the F_{ST} -based population branch statistic (PBS), with selected genes coloured to indicate the population in which the maximum value was achieved.

with frequency $<0.5\%$, and differences in the number of tagging variants between continental groups are less marked.

Among variants in the GWAS catalogue (which have an average frequency of 26.6% in project haplotypes), the number of proxies averages 14.4 in African populations and 30.3–44.4 in other continental groupings (Supplementary Table 10). The potential value of

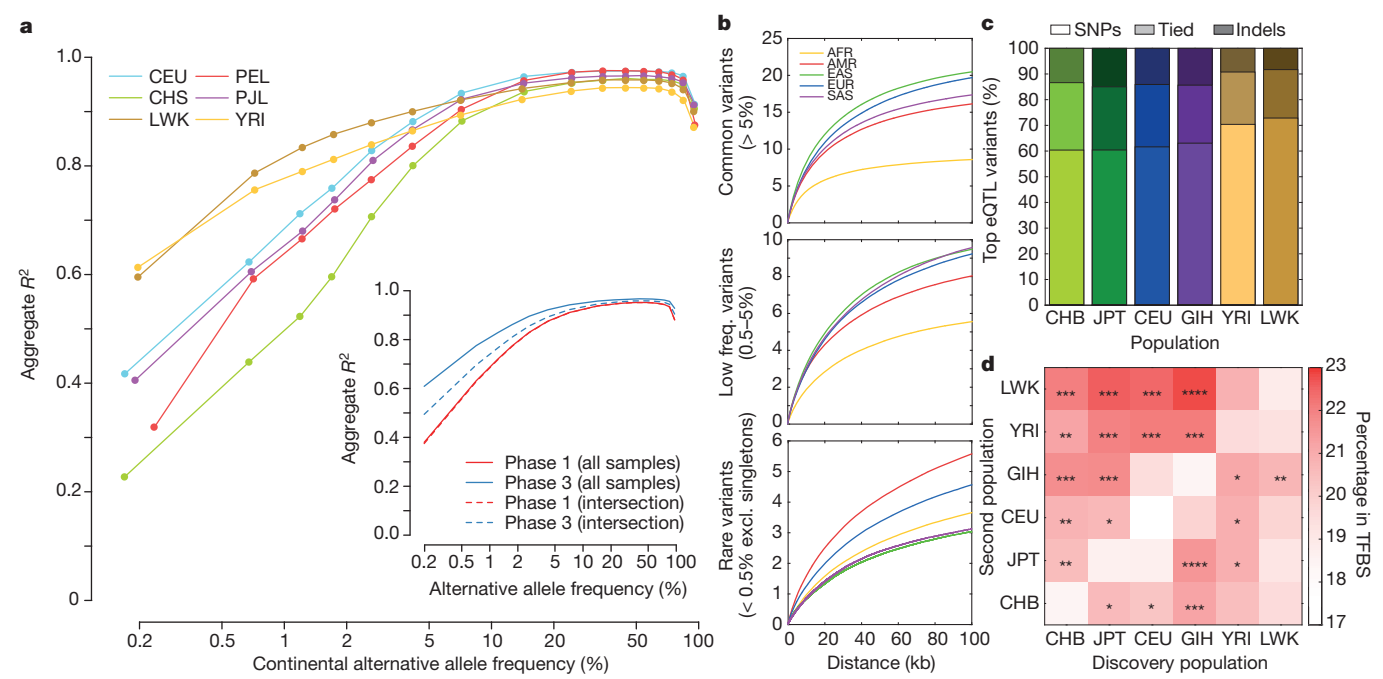


Figure 4 | Imputation and eQTL discovery. **a**, Imputation accuracy as a function of allele frequency for six populations. The insert compares imputation accuracy between phase 3 and phase 1, using all samples (solid lines) and intersecting samples (dashed lines). **b**, The average number of tagging variants ($r^2 > 0.8$) as a function of physical distance for common (top), low frequency (middle), and rare (bottom) variants. **c**, The proportion of top

eQTL variants that are SNPs and indels, as discovered in 69 samples from each population. **d**, The percentage of eQTLs in TFBS, having performed discovery in the first population, and fine mapped by including an additional 69 samples from a second population ($*P < 0.01$, $**P < 0.001$, $***P < 0.0001$, McNemar's test). The diagonal represents the percentage of eQTLs in TFBS using the original discovery sample.

multi-population fine-mapping is illustrated by the observation that the number of proxies shared across all populations is only 8.2 and, furthermore, that 34.9% of GWAS catalogue variants have no proxy shared across all continental groupings.

To further assess prospects for fine-mapping genetic association signals, we performed expression quantitative trait loci (eQTL) discovery at 17,667 genes in 69 samples from each of 6 populations (CEU, CHB, GIH, JPT, LWK, and YRI)³². We identified eQTLs for 3,285 genes at 5% FDR (average 1,265 genes per population). Overall, a typical eQTL signal comprised 67 associated variants, including an indel as one of the top associated variants 26–40% of the time (Fig. 4c). Within each discovery population, 17.5–19.5% of top eQTL variants overlapped annotated transcription factor binding sites (TFBSs), consistent with the idea that a substantial fraction of eQTL polymorphisms are TFBS polymorphisms. Using a meta-analysis approach to combine pairs of populations, the proportion of top eQTL variants overlapping TFBSs increased to 19.2–21.6% (Fig. 4d), consistent with improved localization. Including an African population provided the greatest reduction in the count of associated variants and the greatest increase in overlap between top variants and TFBSs.

Discussion

Over the course of the 1000 Genomes Project there have been substantial advances in sequence data generation, archiving and analysis. Primary sequence data production improved with increased read length and depth, reduced per-base errors, and the introduction of paired-end sequencing. Sequence analysis methods improved with the development of strategies for identifying and filtering poor-quality data, for more accurate mapping of sequence reads (particularly in repetitive regions), for exchanging data between analysis tools and enabling ensemble analyses, and for capturing more diverse types of variants. Importantly, each release has examined larger numbers of individuals, aiding population-based analyses that identify and leverage shared haplotypes during genotyping. Whereas our first analyses produced high-confidence short-variant calls for 80–85% of the reference genome¹, our newest analyses reach ~96% of the genome using the same metrics, although our ability to accurately capture structural variation remains more limited³³. In addition, the evolution of sequencing, analysis and filtering strategies means that our results are not a simple superset of previous analysis. Although the number of characterized variants has more than doubled relative to phase 1, ~2.3 million previously described variants are not included in the current analysis; most missing variants were rare or marked as low quality: 1.6 million had frequency <0.5% and may be missing from our current read set, while the remainder were removed by our filtering processes.

These same technical advances are enabling the application of whole genome sequencing to a variety of medically important samples. Some of these studies already exceed the 1000 Genomes Project in size^{34–36}, but the results described here remain a prime resource for studies of genetic variation for several reasons. First, the 1000 Genomes Project samples provide a broad representation of human genetic variation—in contrast to the bulk of complex disease studies in humans, which primarily study European ancestry samples and which, as we show, fail to capture functionally important variation in other populations. Second, the project analyses incorporate multiple analysis strategies, callsets and variant types. Although such ensemble analyses are cumbersome, they provide a benchmark for what can be achieved and a yardstick against which more practical analysis strategies can be evaluated. Third, project samples and data resulting from them can be shared broadly, enabling sequencing strategies and analysis methods to be compared easily on a benchmark set of samples. Because of the wide availability of the data and samples, these samples have been and will continue to be used for studying many molecular phenotypes. Thus, we predict that the samples will accumulate many

types of data that will allow connections to be drawn between variants and both molecular and disease phenotypes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 May; accepted 20 August 2015.

1. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
2. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
3. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8**, e1002793 (2012).
4. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genet.* **43**, 1193–1201 (2011).
5. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genet.* **44**, 955–959 (2012).
6. Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91**, 1022–1032 (2012).
7. Jung, H., Bleazard, T., Lee, J. & Hong, D. Systematic investigation of cancer-associated somatic point mutations in SNP databases. *Nature Biotechnol.* **31**, 787–789 (2013).
8. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* <http://dx.doi.org/10.1038/nature15394> (this issue).
9. The Haplotype Reference Consortium (<http://www.haplotype-reference-consortium.org/>).
10. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genet.* **46**, 220–224 (2014).
11. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genet.* **47**, 126–131 (2015).
12. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
13. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet.* **10**, e1004528 (2014).
14. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
15. Moltke, I. *et al.* A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
16. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
17. Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**, 1782–1786 (2005).
18. Eiberg, H. *et al.* Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the *HERC2* gene inhibiting *OCA2* expression. *Hum. Genet.* **123**, 177–187 (2008).
19. Mathias, R. A. *et al.* Adaptive evolution of the *FADS* gene cluster within Africa. *PLoS ONE* **7**, e44926 (2012).
20. Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920–924 (2011).
21. Chen, W. *et al.* Genetic variants near *TIMP3* and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proc. Natl Acad. Sci. USA* **107**, 7401–7406 (2010).
22. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
23. Wakefield, J. Commentary: genome-wide significance thresholds via Bayes factors. *Int. J. Epidemiol.* **41**, 286–291 (2012).
24. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nature Rev. Genet.* **15**, 335–346 (2014).
25. Gold, B. *et al.* Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nature Genet.* **38**, 458–462 (2006).
26. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
27. Rivera, A. *et al.* Hypothetical *LOC387715* is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum. Mol. Genet.* **14**, 3227–3236 (2005).
28. Yates, J. R. *et al.* Complement C3 variant and the risk of age-related macular degeneration. *N. Engl. J. Med.* **357**, 553–561 (2007).
29. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genet.* **44**, 1294–1301 (2012).
30. Fritsche, L. G. *et al.* Age-related macular degeneration is associated with an unstable *ARMS2* (*LOC387715*) mRNA. *Nature Genet.* **40**, 892–896 (2008).
31. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
32. Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).
33. Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).

34. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genet.* **47**, 435–444 (2015).
35. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* <http://dx.doi.org/10.1038/nature14962> (2015).
36. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature Genet.* <http://dx.doi.org/10.1038/ng3368> (2015).
37. Delaneau, O. & Marchini, J. The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Commun.* **5**, 3934 (2014).
38. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
39. Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84–91 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the many people who were generous with contributing their samples to the project: the African Caribbean in Barbados; Bengali in Bangladesh; British in England and Scotland; Chinese Dai in Xishuangbanna, China; Colombians in Medellín, Colombia; Esan in Nigeria; Finnish in Finland; Gambian in Western Division – Mandinka; Gujarati Indians in Houston, Texas, USA; Han Chinese in Beijing, China; Iberian populations in Spain; Indian Telugu in the UK; Japanese in Tokyo, Japan; Kinh in Ho Chi Minh City, Vietnam; Luhya in Webuye, Kenya; Mende in Sierra Leone; people with African ancestry in the southwest USA; people with Mexican ancestry in Los Angeles, California, USA; Peruvians in Lima, Peru; Puerto Ricans in Puerto Rico; Punjabi in Lahore, Pakistan; southern Han Chinese; Sri Lankan Tamil in the UK; Toscani in Italy; Utah residents (CEPH) with northern and western European ancestry; and Yoruba in Ibadan, Nigeria. Many thanks to the people who contributed to this project: P. Maul, T. Maul, and C. Foster; Z. Chong, X. Fan, W. Zhou, and T. Chen; N. Sengamalai, S. Ott, L. Sadzewicz, J. Liu, and L. Tallon; L. Merson; O. Folari, D. Asogun, O. Ikponmose, E. Philomena, G. Akpede, S. Okhobgenin, and O. Omoniwa; the staff of the Institute of Lassa Fever Research and Control (ILFRC), Irrua Specialist Teaching Hospital, Irrua, Edo State, Nigeria; A. Schlattl and T. Zichner; S. Lewis, E. Appelbaum, and L. Fulton; A. Yurovsky and I. Padoleau; N. Kaelin and F. Laplace; E. Drury and H. Arbery; A. Naranjo, M. Victoria Parra, and C. Duque; S. Dökel, B. Lenz, and S. Schrunner; S. Bumpstead; and C. Fletcher-Hoppe. Funding for this work was from the Wellcome Trust Core Award 090532/Z/09/Z and Senior Investigator Award 095552/Z/11/Z (P.D.), and grants WT098051 (R.D.), WT095908 and WT109497 (P.F.), WT086084/Z/08/Z and WT100956/Z/13/Z (G.M.), WT097307 (W.K.), WT0855322/Z/08/Z (R.L.), WT090770/Z/09/Z (D.K.), the Wellcome Trust Major Overseas program in Vietnam grant 089276/Z/09/Z (S.D.), the Medical Research Council UK grant G0801823 (J.L.M.), the UK Biotechnology and Biological Sciences Research Council grants BB/I02593X/1 (G.M.) and BB/I021213/1 (A.R.L.), the British Heart Foundation (C.A.A.), the Monument Trust (J.H.), the European Molecular Biology Laboratory (P.F.), the European Research Council grant 617306 (J.L.M.), the Chinese 863 Program 2012AA02A201, the National Basic Research program of China 973 program no. 2011CB809201, 2011CB809202 and 2011CB809203, Natural Science Foundation of China 31161130357, the Shenzhen Municipal Government of China grant ZYC201105170397A (J.W.), the Canadian Institutes of Health Research

Operating grant 136855 and Canada Research Chair (S.G.), Banting Postdoctoral Fellowship from the Canadian Institutes of Health Research (M.K.D.), a Le Fonds de Recherche du Québec-Santé (FRQS) research fellowship (A.H.), Genome Quebec (P.A.), the Ontario Ministry of Research and Innovation – Ontario Institute for Cancer Research Investigator Award (P.A., J.S.), the Quebec Ministry of Economic Development, Innovation, and Exports grant PSR-SIIRI-195 (P.A.), the German Federal Ministry of Education and Research (BMBF) grants 0315428A and 01GS08201 (R.H.), the Max Planck Society (H.L., G.M., R.S.), BMBF-EPITREAT grant 0316190A (R.H., M.L.), the German Research Foundation (Deutsche Forschungsgemeinschaft) Emmy Noether Grant KO4037/1-1 (J.O.K.), the Beatriz de Pinos Program grants 2006 BP-A 10144 and 2009 BP-B 00274 (M.V.), the Spanish National Institute for Health Research grant PRB2 IPT13/0001-ISCIII-SGEFI/FEDER (A.O.), Ewha Womans University (C.L.), the Japan Society for the Promotion of Science Fellowship number PE13075 (N.P.), the Louis Jeantet Foundation (E.T.D.), the Marie Curie Actions Career Integration grant 303772 (C.A.), the Swiss National Science Foundation 31003A_130342 and NCCR “Frontiers in Genetics” (E.T.D.), the University of Geneva (E.T.D., T.L., G.M.), the US National Institutes of Health National Center for Biotechnology Information (S.S.) and grants U54HG3067 (E.S.L.), U54HG3273 and U01HG5211 (R.A.G.), U54HG3079 (R.K.W., E.R.M.), R01HG2898 (S.E.D.), R01HG2385 (E.E.E.), RC2HG5552 and U01HG6513 (G.T.M., G.R.A.), U01HG5214 (A.C.), U01HG5715 (C.D.B.), U01HG5718 (M.G.), U01HG5728 (Y.X.F.), U41HG7635 (R.K.W., E.E.E., P.H.S.), U41HG7497 (C.L., M.A.B., K.C., L.D., E.E.E., M.G., J.O.K., G.T.M., S.A.M., R.E.M., J.L.S., K.Y.), R01HG4960 and R01HG5701 (B.L.B.), R01HG5214 (G.A.), R01HG6855 (S.M.), R01HG7068 (R.E.M.), R01HG7644 (R.D.H.), DP2OD6514 (P.S.), DP5OD9154 (J.K.), R01CA166661 (S.E.D.), R01CA172652 (K.C.), P01GM99568 (S.R.B.), R01GM59290 (L.B.J., M.A.B.), R01GM104390 (L.B.J., M.Y.Y.), T32GM7790 (C.D.B., A.R.M.), P01GM99568 (S.R.B.), R01HL87699 and R01HL104608 (K.C.B.), T32HL94284 (J.L.R.F.), and contracts HHSN268201100040C (A.M.R.) and HHSN272201000025C (P.S.), Harvard Medical School Eleanor and Miles Shore Fellowship (K.L.), Lundbeck Foundation Grant R170-2014-1039 (K.L.), NIJ Grant 2014-DN-BX-K089 (Y.E.), the Mary Beryl Patch Turnbull Scholar Program (K.C.B.), NSF Graduate Research Fellowship DGE-1147470 (G.D.P.), the Simons Foundation SFARI award SF51 (M.W.), and a Sloan Foundation Fellowship (R.D.H.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author Contributions Details of author contributions can be found in the author list.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.A. (adam.auton@gmail.com) or G.R.A. (goncalo@umich.edu).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

The 1000 Genomes Project Consortium (Participants are arranged by project role, then by institution alphabetically, and finally alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.)

Corresponding authors Adam Auton¹, Gonçalo R. Abecasis²

Steering committee: David M. Altshuler³ (Co-Chair), Richard M. Durbin⁴ (Co-Chair), Gonçalo R. Abecasis², David R. Bentley⁵, Aravinda Chakravarti⁶, Andrew G. Clark⁷, Peter Donnelly^{8,9}, Evan E. Eichler^{10,11}, Paul Flicek¹², Stacey B. Gabriel¹³, Richard A. Gibbs¹⁴, Eric D. Green¹⁵, Matthew E. Hurles⁴, Bartha M. Knoppers¹⁶, Jan O. Korbel^{12,17}, Eric S. Lander¹³, Charles Lee^{18,19}, Hans Lehrach^{20,21}, Elaine R. Mardis²², Gabor T. Marth²³, Gil A. McVean^{8,9}, Deborah A. Nickerson¹⁰, Jeanette P. Schmidt²⁴, Stephen T. Sherry²⁵, Jun Wang^{26,27,28,29,30}, Richard K. Wilson²²

Production group: **Baylor College of Medicine** Richard A. Gibbs¹⁴ (Principal Investigator), Eric Boerwinkle¹⁴, Harsha Doddapaneni¹⁴, Yi Han¹⁴, Viktoriya Korchina¹⁴, Christie Kovar¹⁴, Sandra Lee¹⁴, Donna Muzny¹⁴, Jeffrey G. Reid¹⁴, Yiming Zhu¹⁴; **BGI-Shenzhen** Jun Wang^{26,27,28,29,30} (Principal Investigator), Yuqi Chang²⁶, Qiang Feng^{26,27}, Xiaodong Fang^{26,27}, Xiaosen Guo^{26,27}, Min Jian^{26,27}, Hui Jiang^{26,27}, Xin Jin²⁶, Tianming Lan²⁶, Guoqing Li²⁶, Jingxiang Li²⁶, Yingrui Li²⁶, Shengmao Liu²⁶, Xiao Liu^{26,27}, Yao Lu²⁶, Xuodi Ma²⁶, Meifang Tang²⁶, Bo Wang²⁶, Guangbiao Wang²⁶, Honglong Wu²⁶, Renhua Wu²⁶, Xun Xu²⁶, Ye Yin²⁶, Dandan Zhang²⁶, Wenwei Zhang²⁶, Jiao Zhao²⁶, Meiru Zhao²⁶, Xiaole Zheng²⁶; **Broad Institute of MIT and Harvard** Eric S. Lander¹³ (Principal Investigator), David M. Altshuler³, Stacey B. Gabriel¹³ (Co-Chair), Namrata Gupta¹³; **Coriell Institute for Medical Research** Neda Gharani³¹, Lorraine H. Toji³¹, Norman P. Gerry³¹, Alissa M. Resch³¹; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Jonathan Barker¹², Laura Clarke¹², Laurent Gil¹², Sarah E. Hunt¹², Gavin Kelman¹², Eugene Kulesha¹², Rasko Leinonen¹², William M. McLaren¹², Rajesh Radhakrishnan¹², Asier Roa¹², Dmitry Smirnov¹², Richard E. Smith¹², Ian Streeter¹², Anja Thormann¹², Iliana Toneva¹², Brendan Vaughan¹², Xiangqun Zheng-Bradley¹²; **illumina** David R. Bentley⁵ (Principal Investigator), Russell Crocock², Sean Humphray², Terena James², Zoya Kingsbury²; **Max Planck Institute for Molecular Genetics** Hans Lehrach^{20,21} (Principal Investigator), Ralf Sudbrak³² (Project Leader), Marcus W. Albrecht³³, Vyacheslav S. Amstislavskiy²⁰, Tatiana A. Borodina³³, Matthias Lienhard²⁰, Florian Mertes²⁰, Marc Sultan²⁰, Bernd Timmermann²⁰, Marie-Laure Yaspo²⁰; **McDonnell Genome Institute at Washington University** Elaine R. Mardis²² (Co-Principal Investigator) (Co-Chair), Richard K. Wilson²² (Co-Principal Investigator), Lucinda Fulton²², Robert Fulton²²; **US National Institutes of Health** Stephen T. Sherry²⁵ (Principal Investigator), Victor Ananiev²⁵, Zinaida Belaia²⁵, Dimitriy Beloslyudtsev²⁵, Nathan Bouk²⁵, Chao Chen²⁵, Deanna Church³⁴, Robert Cohen²⁵, Charles Cook²⁵, John Garner²⁵, Timothy Hefferon²⁵, Mikhail Kimelman²⁵, Chunlei Liu²⁵, John Lopez²⁵, Peter Meric²⁵, Chris O'Sullivan³⁵, Yuri Ostapchuk²⁵, Lon Phan²⁵, Sergiy Ponomarev²⁵, Valerie Schneider²⁵, Eugene Shekhtman²⁵, Karl Sirotkin²⁵, Douglas Slotta²⁵, Hua Zhang²⁵; **University of Oxford** Gil A. McVean^{8,9} (Principal Investigator); **Wellcome Trust Sanger Institute** Richard M. Durbin⁴ (Principal Investigator), Senduran Balasubramanian⁴, John Burton⁴, Petr Danecek⁴, Thomas M. Keane⁴, Anja Kolb-Kokocinski⁴, Shane McCarthy⁴, James Stalker⁴, Michael Quail⁴

Analysis group: **Affymetrix** Jeanette P. Schmidt²⁴ (Principal Investigator), Christopher J. Davies²⁴, Jeremy Gollub²⁴, Teresa Webster²⁴, Brant Wong²⁴, Yiping Zhan²⁴; **Albert Einstein College of Medicine** Adam Auton¹ (Principal Investigator), Christopher L. Campbell¹, Yu Kong¹, Anthony Marcketta¹; **Baylor College of Medicine** Richard A. Gibbs¹⁴ (Principal Investigator), Fuli Yu¹⁴ (Project Leader), Lilian Antunes¹⁴, Matthew Bainbridge¹⁴, Donna Muzny¹⁴, Aniko Sabo¹⁴, Zhuoyi Huang¹⁴; **BGI-Shenzhen** Jun Wang^{26,27,28,29,30} (Principal Investigator), Lachlan J. M. Coin²⁶, Lin Fang^{26,27}, Xiaosen Guo²⁶, Xin Jin²⁶, Guoqing Li²⁶, Qibin Li²⁶, Yingrui Li²⁶, Zhenyu Li²⁶, Haoxiang Lin²⁶, Binghang Liu²⁶, Ruibang Luo²⁶, Haojing Shao²⁶, Yinlong Xie²⁶, Chen Ye²⁶, Chang Yu²⁶, Fan Zhang²⁶, Hancheng Zheng²⁶, Hongmei Zhu²⁶; **Bilkent University** Can Alkan³⁶, Elif Dal³⁶, Fatma Kahveci³⁶; **Boston College** Gabor T. Marth²³ (Principal Investigator), Erik P. Garrison⁴ (Project Lead), Deniz Kural³⁷, Wan-Ping Lee³⁷, Wen Fung Leong³⁸, Michael Stromberg³⁹, Alistair N. Ward²³, Jiantao Wu³⁹, Mengyao Zhang⁴⁰; **Broad Institute of MIT and Harvard** Mark J. Daly¹³ (Principal Investigator), Mark A. DePristo⁴¹ (Project Leader), Robert E. Handsaker^{13,40} (Project Leader), David M. Altshuler³, Eric Banks¹³, Gaurav Bhatia¹³, Guillermo del Angel¹³, Stacey B. Gabriel¹³, Giulio Genovesi¹³, Namrata Gupta¹³, Heng Li¹³, Seva Kashin^{13,40}, Eric S. Lander¹³, Steven A. McCarroll^{13,40}, James C. Nemes¹³, Ryan E. Poplin¹³; **Cold Spring Harbor Laboratory** Seungtae C. Yoon⁴² (Principal Investigator), Jayon Lihm⁴², Vladimir Makarov⁴³; **Cornell University** Andrew G. Clark⁷ (Principal Investigator), Srikanth Gottipati⁴⁴, Alon Keinan⁷, Juan L. Rodriguez-Flores⁴⁵; **European Molecular Biology Laboratory** Jan O. Korbel^{12,17} (Principal Investigator), Tobias Rausch^{17,46} (Project Leader), Markus H. Fritz⁴⁶, Adrian M. Stütz¹⁷; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Kathryn Beal¹², Laura Clarke¹², Avik Datta¹², Javier Herrero⁴⁷, William M. McLaren¹², Graham R. S. Ritchie¹², Richard E. Smith¹², Daniel Zerbinio¹², Xiangqun Zheng-Bradley¹²; **Harvard University** Pardis C. Sabeti^{13,48} (Principal Investigator), Ilya Shlyakhter^{13,48}, Stephen F. Schaffner^{13,48}, Joseph Vitti^{13,49}; **Human Gene Mutation Database** David N. Cooper⁵⁰ (Principal Investigator), Edward V. Ball⁵⁰, Peter D. Stenson⁵⁰; **illumina** David R. Bentley⁵ (Principal Investigator), Bret Barnes³⁹, Markus Bauer⁵, R. Keira Cheetham⁵, Anthony Cox⁵, Michael Eberle⁵, Sean Humphray⁵, Scott Kahn⁵, Lisa Murray⁵, John Peden⁵, Richard Shaw⁵; **Icahn School of Medicine at Mount Sinai** Eimear E. Kenny⁵¹ (Principal Investigator); **Louisiana State University** Mark A. Batzer⁵² (Principal Investigator), Miriam K. Konkel⁵², Jerilyn A. Walker⁵²; **Massachusetts General Hospital** Daniel G. MacArthur⁵³ (Principal Investigator), Monkol Lek⁵³; **Max Planck Institute for Molecular Genetics** Ralf Sudbrak³² (Project Leader), Vyacheslav S. Amstislavskiy²⁰, Ralf Herwig²⁰; **McDonnell Genome Institute at Washington University** Elaine R. Mardis²² (Co-Principal Investigator), Li Ding²², Daniel C. Koboldt²², David Larson²², Kai

Ye²²; **McGill University** Simon Gravel⁵⁴; **National Eye Institute, NIH** Anand Swaroop⁵⁵, Emily Chew⁵⁵; **New York Genome Center** Tuuli Lappalainen^{56,57} (Principal Investigator), Yaniv Erlich^{56,58} (Principal Investigator), Melissa Gymrek^{13,56,59,60}, Thomas Frederick Willems⁶¹; **Ontario Institute for Cancer Research** Jared T. Simpson⁶²; **Pennsylvania State University** Mark D. Shriver⁶³ (Principal Investigator); **Rutgers Cancer Institute of New Jersey** Jeffrey A. Rosenfeld⁶⁴ (Principal Investigator); **Stanford University** Carlos D. Bustamante⁶⁵ (Principal Investigator), Stephen B. Montgomery⁶⁶ (Principal Investigator), Francisco M. De La Vega⁶⁵ (Principal Investigator), Jake K. Byrnes⁶⁷, Andrew W. Carroll⁶⁸, Marianne K. DeGorter⁶⁶, Phil Lacroute⁶⁵, Brian K. Maples⁶⁵, Alicia R. Martin⁶⁵, Andres Moreno-Estrada^{65,69}, Suyash S. Shringarpure⁶⁵, Fouad Zakharia⁶⁵; **Tel-Aviv University** Eran Halperin^{70,71,72} (Principal Investigator), Yael Baran⁷⁰; **The Jackson Laboratory for Genomic Medicine** Charles Lee^{18,19} (Principal Investigator), Eliza Cerveira¹⁸, Jaeho Hwang¹⁸, Ankit Malhotra¹⁸ (Co-Project Lead), Dariusz Plewczynski¹⁸, Kamen Radew¹⁸, Mallory Romanovitch¹⁸, Chengsheng Zhang¹⁸ (Co-Project Lead); **Thermo Fisher Scientific** Fiona C. L. Hyland⁷³; **Translational Genomics Research Institute** David W. Craig⁷⁴ (Principal Investigator), Alexis Christoforides⁷⁴, Nils Homer⁷⁵, Tyler Izatt⁷⁴, Ahmet A. Kurdoglu⁷⁴, Shripad A. Sinari⁷⁴, Kevin Squire⁷⁶; **US National Institutes of Health** Stephen T. Sherry²⁵ (Principal Investigator), Chunlin Xiao²⁵; **University of California, San Diego** Jonathan Sebat^{77,78} (Principal Investigator), Danny Antaki⁷⁷, Madhusudan Gujral⁷⁷, Amina Noor⁷⁷, Kenny Ye⁷⁹; **University of California, San Francisco** Esteban G. Burchard⁸⁰ (Principal Investigator), Ryan D. Hernandez^{80,81,82} (Principal Investigator), Christopher R. Gignoux⁸⁰; **University of California, Santa Cruz** David Haussler^{83,84} (Principal Investigator), Sol J. Katzman⁸³, W. James Kent⁸³; **University of Chicago** Bryan Howie⁸⁵; **University College London** Andres Ruiz-Linares⁸⁶ (Principal Investigator); **University of Geneva** Emmanouil T. Dermitakis^{87,88,89} (Principal Investigator); **University of Maryland School of Medicine** Scott E. Devine⁹⁰ (Principal Investigator); **University of Michigan** Gonçalo R. Abecasis² (Principal Investigator) (Co-Chair), Hyun Min Kang² (Project Leader), Jeffrey M. Kidd^{91,92} (Principal Investigator), Tom Blackwell², Sean Caron², Wei Chen⁹³, Sarah Emery⁹², Lars Fritsche², Christian Fuchsberger², Goo Jun^{2,94}, Bingshan Li⁹⁵, Robert Lyons⁹⁶, Chris Scheller², Carlo Sidore^{2,97,98}, Shiya Song⁹¹, Elzbieta Sliwerska⁹², Daniel Taliun², Adrian Tan², Ryan Welch², Mary Kate Wing², Xiaowei Zhan⁹⁹; **University of Montréal** Philip Awadalla^{62,100} (Principal Investigator), Alan Hodgkinson¹⁰⁰; **University of North Carolina at Chapel Hill** Yun Li¹⁰¹; **University of North Carolina at Charlotte** Xinghua Shi¹⁰² (Principal Investigator), Andrew Quitadamo¹⁰²; **University of Oxford** Gordon Lunter⁵ (Principal Investigator), Gil A. McVean^{8,9} (Principal Investigator) (Co-Chair), Jonathan L. Marchini^{8,9} (Principal Investigator), Simon Myers^{8,9} (Principal Investigator), Claire Churchhouse⁵, Olivier Delaneau^{9,87}, Anjali Gupta-Hinch⁸, Warren Kretschmar⁸, Zamin Iqbal⁸, Iain Mathieson⁸, Androniki Menelaou^{9,103}, Andy Rimmer⁸⁷, Dionysia K. Xifara^{8,9}; **University of Puerto Rico** Taras K. Oleksyk¹⁰⁴ (Principal Investigator); **University of Texas Health Sciences Center at Houston** Yunxin Fu⁹⁴ (Principal Investigator), Xiaoming Liu⁹⁴, Momiao Xiong⁹⁴; **University of Utah** Lynn Jorde¹⁰⁵ (Principal Investigator), David Witherpoon¹⁰⁵, Jinchuan Xing¹⁰⁶; **University of Washington** Evan E. Eichler^{10,11} (Principal Investigator), Brian L. Browning¹⁰⁷ (Principal Investigator), Sharon R. Browning¹⁰⁸ (Principal Investigator), Fereydoon Hormozdizadeh¹⁰, Peter H. Sudmant¹⁰; **Weill Cornell Medical College**, Ekta Khurana¹⁰⁹ (Principal Investigator); **Wellcome Trust Sanger Institute** Richard M. Durbin⁴ (Principal Investigator), Matthew E. Hurles⁴ (Principal Investigator), Chris Tyler-Smith⁴ (Principal Investigator), Cornelis A. Albers^{110,111}, Qasim Ayub⁴, Senduran Balasubramanian⁴, Yuan Chen⁴, Vincenza Colonna^{4,112}, Petr Danecek⁴, Luke Josins⁸, Thomas M. Keane⁴, Shane McCarthy⁴, Klaudia Walter⁴, Yali Xue⁴; **Yale University** Mark B. Gerstein^{113,114,115} (Principal Investigator), Alexey Abyzov¹¹⁶, Suganthi Balasubramanian¹¹⁵, Jieming Chen¹¹³, Declan Clarke¹¹⁷, Yao Fu¹¹³, Arif O. Harmanci¹¹³, Mike Jin¹¹⁵, Donghoon Lee¹¹³, Jeremy Liu¹¹⁵, Ximeng Jasmine Mu^{113,113}, Jing Zhang^{113,115}, Yan Zhang^{113,115}

Structural variation group: **BGI-Shenzhen** Yingrui Li²⁶, Ruibang Luo²⁶, Hongmei Zhu²⁶; **Bilkent University** Can Alkan³⁶, Elif Dal³⁶, Fatma Kahveci³⁶; **Boston College** Gabor T. Marth²³ (Principal Investigator), Erik P. Garrison⁴, Deniz Kural³⁷, Wan-Ping Lee³⁷, Alistair N. Ward²³, Jiantao Wu²³, Mengyao Zhang²³; **Broad Institute of MIT and Harvard** Steven A. McCarroll^{13,40} (Principal Investigator), Robert E. Handsaker^{13,40} (Project Leader), David M. Altshuler³, Eric Banks¹³, Guillermo del Angel¹³, Giulio Genovesi¹³, Chris Hart¹³, Heng Li¹³, Seva Kashin^{13,40}, James C. Nemes¹³, Khalid Shakir¹³; **Cold Spring Harbor Laboratory** Seungtae C. Yoon⁴² (Principal Investigator), Jayon Lihm⁴², Vladimir Makarov⁴³; **Cornell University** Jeremiah Degenhardt⁷; **European Molecular Biology Laboratory** Jan O. Korbel^{12,17} (Principal Investigator) (Co-Chair), Markus H. Fritz⁴⁶, Sascha Meiers¹⁷, Benjamin Raeder¹⁷, Tobias Rausch^{17,46}, Adrian M. Stütz¹⁷; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Francesco Paolo Casale¹², Laura Clarke¹², Richard E. Smith¹², Oliver Stegle¹², Xiangqun Zheng-Bradley¹²; **illumina** David R. Bentley⁵ (Principal Investigator), Bret Barnes³⁹, R. Keira Cheetham⁵, Michael Eberle⁵, Sean Humphray⁵, Scott Kahn⁵, Lisa Murray⁵, Richard Shaw⁵; **Leiden University Medical Center** Eric-Wubbo Lameijer¹¹⁸; **Louisiana State University** Mark A. Batzer⁵² (Principal Investigator), Miriam K. Konkel⁵², Jerilyn A. Walker⁵²; **McDonnell Genome Institute at Washington University** Li Ding²² (Principal Investigator), Ira Hall²², Kai Ye²²; **Stanford University** Phil Lacroute⁶⁵; **The Jackson Laboratory for Genomic Medicine** Charles Lee^{18,19} (Principal Investigator) (Co-Chair), Eliza Cerveira¹⁸, Ankit Malhotra¹⁸, Jaeho Hwang¹⁸, Dariusz Plewczynski¹⁸, Kamen Radew¹⁸, Mallory Romanovitch¹⁸, Chengsheng Zhang¹⁸; **Translational Genomics Research Institute** David W. Craig⁷⁴ (Principal Investigator), Nils Homer⁷⁵; **US National Institutes of Health** Deanna Church³⁴, Chunlin Xiao²⁵; **University of California, San Diego** Jonathan Sebat⁷⁷ (Principal Investigator), Danny Antaki⁷⁷, Vineet Bafna¹¹⁹, Jacob Michaelson¹²⁰, Kenny Ye⁷⁹; **University of Maryland School of Medicine** Scott E. Devine⁹⁰ (Principal Investigator), Eugene J. Gardner⁹⁰ (Project Leader); **University of Michigan** Gonçalo R. Abecasis² (Principal Investigator), Jeffrey M. Kidd^{91,92} (Principal Investigator), Ryan E. Mills^{91,92} (Principal Investigator), Gargi

Dayama^{91,92}, Sarah Emery⁹², Goo Jun^{2,94}; **University of North Carolina at Charlotte** Xinghua Shi¹⁰² (Principal Investigator), Andrew Quitadamo¹⁰²; **University of Oxford** Gerton Lunter⁸ (Principal Investigator), Gil A. McVean^{8,9} (Principal Investigator); **University of Texas MD Anderson Cancer Center** Ken Chen¹²¹ (Principal Investigator), Xian Fan¹²¹, Zechen Chong¹²¹, Tenghui Chen¹²¹; **University of Utah** David Witherspoon¹⁰⁵, Jinchuan Xing¹⁰⁶; **University of Washington** Evan E. Eichler^{10,11} (Principal Investigator) (Co-Chair), Mark J. Chaisson¹⁰, Fereydoon Hormozdian¹⁰, John Huddleston^{10,11}, Maika Malig¹⁰, Bradley J. Nelson¹⁰, Peter H. Sudmant¹⁰; **Vanderbilt University School of Medicine** Nicholas F. Parrish⁹⁵; **Weill Cornell Medical College** Ekta Khurana¹⁰⁹ (Principal Investigator); **Wellcome Trust Sanger Institute** Matthew E. Hurles⁴ (Principal Investigator), Ben Blackburne⁴, Sarah J. Lindsay⁴, Zemin Ning⁴, Klaudia Walter⁴, Yujun Zhang⁴; **Yale University** Mark B. Gerstein^{113,114,115} (Principal Investigator), Alexej Abyzov¹⁶, Jieming Chen¹¹³, Declan Clarke¹⁷, Hugo Lam¹²², Xinnmeng Jasmine Mu^{13,113}, Cristina Sisu¹¹³, Jing Zhang^{113,115}, Yan Zhang^{113,115}

Exome group: Baylor College of Medicine Richard A. Gibbs¹⁴ (Principal Investigator) (Co-Chair), Fuli Yu¹⁴ (Project Leader), Matthew Bainbridge¹⁴, Danny Challis¹⁴, Uday S. Evans¹⁴, Christie Kovar¹⁴, James Lu¹⁴, Donna Muzny¹⁴, Uma Nagaswamy¹⁴, Jeffrey G. Reid¹⁴, Aniko Sabo¹⁴, Jin Yu¹⁴; **BGI-Shenzhen** Xiaosen Guo^{26,27}, Wangshen Li²⁶, Yingrui Li²⁶, Renhua Wu²⁶; **Boston College** Gabor T. Marth²³ (Principal Investigator) (Co-Chair), Erik P. Garrison⁴, Wen Fung Leong²³, Alistair N. Ward²³; **Broad Institute of MIT and Harvard** Guillermo del Angel¹³, Mark A. DePristo⁴¹, Stacey B. Gabriel¹³, Namrata Gupta¹³, Chris Hartl¹³, Ryan E. Poplin¹³; **Cornell University** Andrew G. Clark⁷ (Principal Investigator), Juan L. Rodriguez-Flores⁴⁵; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Laura Clarke¹², Richard E. Smith¹², Xiangqun Zheng-Bradley¹²; **Massachusetts General Hospital** Daniel G. MacArthur⁵³ (Principal Investigator); **McDonnell Genome Institute at Washington University** Elaine R. Mardis²² (Principal Investigator); Robert Fulton²², Daniel C. Koboldt²²; **McGill University** Simon Gravel⁵⁴; **Stanford University** Carlos D. Bustamante⁶⁵ (Principal Investigator); **Translational Genomics Research Institute** David W. Craig⁷⁴ (Principal Investigator), Alexis Christoforides⁷⁴, Nils Homer⁷⁵, Tyler Izatt⁷⁴; **US National Institutes of Health** Stephen T. Sherry²⁵ (Principal Investigator), Chunlin Xiao²⁵; **University of Geneva** Emmanouil T. Dermizakis^{87,88,89} (Principal Investigator); **University of Michigan** Gonçalo R. Abecasis² (Principal Investigator), Hyun Min Kang²; **University of Oxford** Gil A. McVean^{8,9} (Principal Investigator); **Yale University** Mark B. Gerstein^{113,114,115} (Principal Investigator), Suganthi Balasubramanian¹¹⁵, Lukas Habegger¹¹³

Functional interpretation group: Cornell University Haiyuan Yu⁴⁴ (Principal Investigator); **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Laura Clarke¹², Fiona Cunningham¹², Ian Dunham¹², Daniel Zerbino¹², Xiangqun Zheng-Bradley¹²; **Harvard University** Kasper Lage^{13,123} (Principal Investigator), Jakob Berg Jaspersen^{13,123,124}, Heiko Horn^{13,123}; **Stanford University** Stephen B. Montgomery⁶⁶ (Principal Investigator), Marianne K. DeGorter⁶⁶; **Weill Cornell Medical College**, Ekta Khurana¹⁰⁹ (Principal Investigator); **Wellcome Trust Sanger Institute** Chris Tyler-Smith⁴ (Principal Investigator) (Co-Chair), Yuan Chen⁴, Vincenza Colonna^{4,112}, Yali Xue⁴; **Yale University** Mark B. Gerstein^{113,114,115} (Principal Investigator) (Co-Chair), Suganthi Balasubramanian¹¹⁵, Yao Fu¹¹³, Donghoon Kim¹¹⁵

Chromosome Y group: Albert Einstein College of Medicine Adam Auton¹ (Principal Investigator), Anthony Marchetta¹; **American Museum of Natural History** Rob Desalle¹²⁵, Apurva Narechania¹²⁶; **Arizona State University** Melissa A. Wilson Sayres¹²⁷; **Boston College** Erik P. Garrison⁴; **Broad Institute of MIT and Harvard** Robert E. Handsaker^{13,40}, Seva Kashin^{13,40}, Steven A. McCarroll^{13,40}; **Cornell University**: Juan L. Rodriguez-Flores⁴⁵; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Laura Clarke¹², Xiangqun Zheng-Bradley¹²; **New York Genome Center** Yaniv Erlich^{56,58}, Melissa Gymrek^{13,56,59,60}, Thomas Frederick Willems⁶¹; **Stanford University** Carlos D. Bustamante⁶⁵ (Principal Investigator) (Co-Chair), Fernando L. Mendez⁶⁵, G. David Poznik¹²⁸, Peter A. Underhill⁶⁵; **The Jackson Laboratory for Genomic Medicine** Charles Lee^{18,19}, Eliza Cerveira¹⁸, Ankit Malhotra¹⁸, Mallory Romanovitch¹⁸, Chengsheng Zhang¹⁸; **University of Michigan** Gonçalo R. Abecasis² (Principal Investigator); **University of Queensland** Lachlan Coin¹²⁹ (Principal Investigator), Haojing Shao¹²⁹; **Virginia Bioinformatics Institute** David Mittelman¹³⁰; **Wellcome Trust Sanger Institute** Chris Tyler-Smith⁴ (Principal Investigator) (Co-Chair), Qasim Ayub⁴, Ruby Banerjee⁴, Maria Cerezo⁴, Yuan Chen⁴, Thomas W. Fitzgerald⁴, Sandra Louzada⁴, Andrea Massai⁴, Shane McCarthy⁴, Graham R. Ritchie⁴, Yali Xue⁴, Fengtang Yang⁴

Data coordination center group: Baylor College of Medicine Richard A. Gibbs¹⁴ (Principal Investigator), Christie Kovar¹⁴, Divya Kalra¹⁴, Walker Hale¹⁴, Donna Muzny¹⁴, Jeffrey G. Reid¹⁴; **BGI-Shenzhen** Jun Wang^{26,27,28,29,30} (Principal Investigator), Xu Dan²⁶, Xiaosen Guo^{26,27}, Guoqing Li²⁶, Yingrui Li²⁶, Chen Ye²⁶, Xiaole Zheng²⁶; **Broad Institute of MIT and Harvard** David M. Altschuler³; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator) (Co-Chair), Laura Clarke¹² (Project Lead), Xiangqun Zheng-Bradley¹²; **Illumina** David R. Bentley³ (Principal Investigator), Anthony Cox³, Sean Humphray³, Scott Kahn³⁹; **Max Planck Institute for Molecular Genetics** Ralf Sudbrak³² (Project Lead), Marcus W. Albrecht³³, Matthias Lienhard²⁰; **McDonnell Genome Institute at Washington University** David Larson²²; **Translational Genomics Research Institute** David W. Craig⁷⁴ (Principal Investigator), Tyler Izatt⁷⁴, Ahmet A. Kurdoglu⁷⁴; **US National Institutes of Health** Stephen T. Sherry²⁵ (Principal Investigator) (Co-Chair), Chunlin Xiao²⁵; **University of California, Santa Cruz** David Haussler^{33,84} (Principal Investigator); **University of Michigan** Gonçalo R. Abecasis² (Principal Investigator); **University of Oxford** Gil A. McVean^{8,9} (Principal Investigator); **Wellcome Trust Sanger Institute** Richard M. Durbin⁴ (Principal Investigator), Senduran Balasubramanian⁴, Thomas M. Keane⁴, Shane McCarthy⁴, James Stalker⁴

Samples and ELSI group: Aravinda Chakravarti⁶ (Co-Chair), Bartha M. Knoppers¹⁶ (Co-Chair), Gonçalo R. Abecasis², Kathleen C. Barnes¹³¹, Christine Beiswanger³¹, Esteban G. Burchard⁸⁰, Carlos D. Bustamante⁶⁵, Hongyu Cai²⁶, Hongzhi Cao^{26,27}, Richard M. Durbin⁴, Norman P. Gerry³¹, Neda Gharani³¹, Richard A. Gibbs¹⁴, Christopher R. Gignoux⁸⁰, Simon Gravel⁵⁴, Brenna Henn¹³², Danielle Jones¹⁴, Lynn Jorde¹⁰⁵, Jane S. Kaye¹³³, Alon Keinan⁷, Alastair Kent¹³⁴, Angeliki Kerasidou¹³⁵, Yingrui Li²⁶, Rasika Mathias¹³⁶, Gil A. McVean^{8,9}, Andres Moreno-Estrada^{65,69}, Pilar N. Ossorio^{137,138}, Michael Parker¹³⁵, Alissa M. Resch³¹, Charles N. Rotimi¹³⁹, Charmaine D. Royal¹⁴⁰, Karla Sandoval⁶⁵, Yeyang Su²⁶, Ralf Sudbrak³², Zhongming Tian²⁶, Sarah Tishkoff¹⁴¹, Lorraine H. Toji³¹, Chris Tyler-Smith⁴, Marc Via¹⁴², Yuhong Wang²⁶, Huanming Yang²⁶, Ling Yang²⁶, Jiayong Zhu²⁶

Sample collection: British from England and Scotland (GBR) Walter Bodmer¹⁴³; **Colombians in Medellín, Colombia (CLM)** Gabriel Bedoya¹⁴⁴, Andres Ruiz-Linares⁸⁶; **Han Chinese South (CHS)** Zhiming Cai²⁶, Yang Gao¹⁴⁵, Jiayou Chu¹⁴⁶; **Finnish in Finland (FIN)** Leena Peltonen[†]; **Iberian Populations in Spain (IBS)** Andres Garcia-Montero¹⁴⁷, Alberto Orfao¹⁴⁷; **Puerto Ricans in Puerto Rico (PUR)** Julie Dutil¹⁴⁸, Juan C. Martinez-Cruzado¹⁰⁴, Taras K. Oleksyk¹⁰⁴; **African Caribbean in Barbados (ACB)** Kathleen C. Barnes¹³¹, Rasika A. Mathias¹³⁶, Anselm Hennis^{149,150}, Harold Watson¹⁵⁰, Colin McKenzie¹⁵¹; **Bengali in Bangladesh (BEB)** Firdausi Qadri¹⁵², Regina LaRocque¹⁵², Pardis C. Sabeti^{13,48}; **Chinese Dai in Xishuangbanna, China (CDX)** Jiayong Zhu²⁶, Xiaoyan Deng¹⁵³; **Esan in Nigeria (ESN)** Pardis C. Sabeti^{13,48}, Danny Asogun¹⁵⁴, Onikepe Folarin¹⁵⁵, Christian Happi^{155,156}, Omonwumi Omonifa^{155,156}, Matt Strelau^{13,48}, Ridhi Tariyal^{13,48}; **Gambian in Western Division – Mandinka (GWD)** Muminatou Jallow^{8,157}, Fatoumatta Sisay Joor^{8,157}, Tumani Corrah^{8,157}, Kirk Rockett^{8,157}, Dominic Kwiatkowski^{8,157}; **Indian Telugu in the UK (ITU)** and **Sri Lankan Tamil in the UK (STU)** Jaspal Kooner¹⁵⁸; **Kinh in Ho Chi Minh City, Vietnam (KHV)** Trần Tịnh Hiền¹⁵⁹, Sarah J. Dunstan^{159,160}, Nguyen Thuy Hang¹⁵⁹; **Mende in Sierra Leone (MSL)** Richard Fonniet¹⁶¹, Robert Garry¹⁶², Lansana Kanneh¹⁶¹, Lina Moses¹⁶², Pardis C. Sabeti^{13,48}, John Schieffelin¹⁶², Donald S. Grant^{161,162}; **Peruvian in Lima, Peru (PEL)** Carla Gallo¹⁶³, Giovanni Poletti¹⁶³; **Punjabi in Lahore, Pakistan (PJL)** Danish Saleheen^{164,165}, Asif Rasheed¹⁶⁴

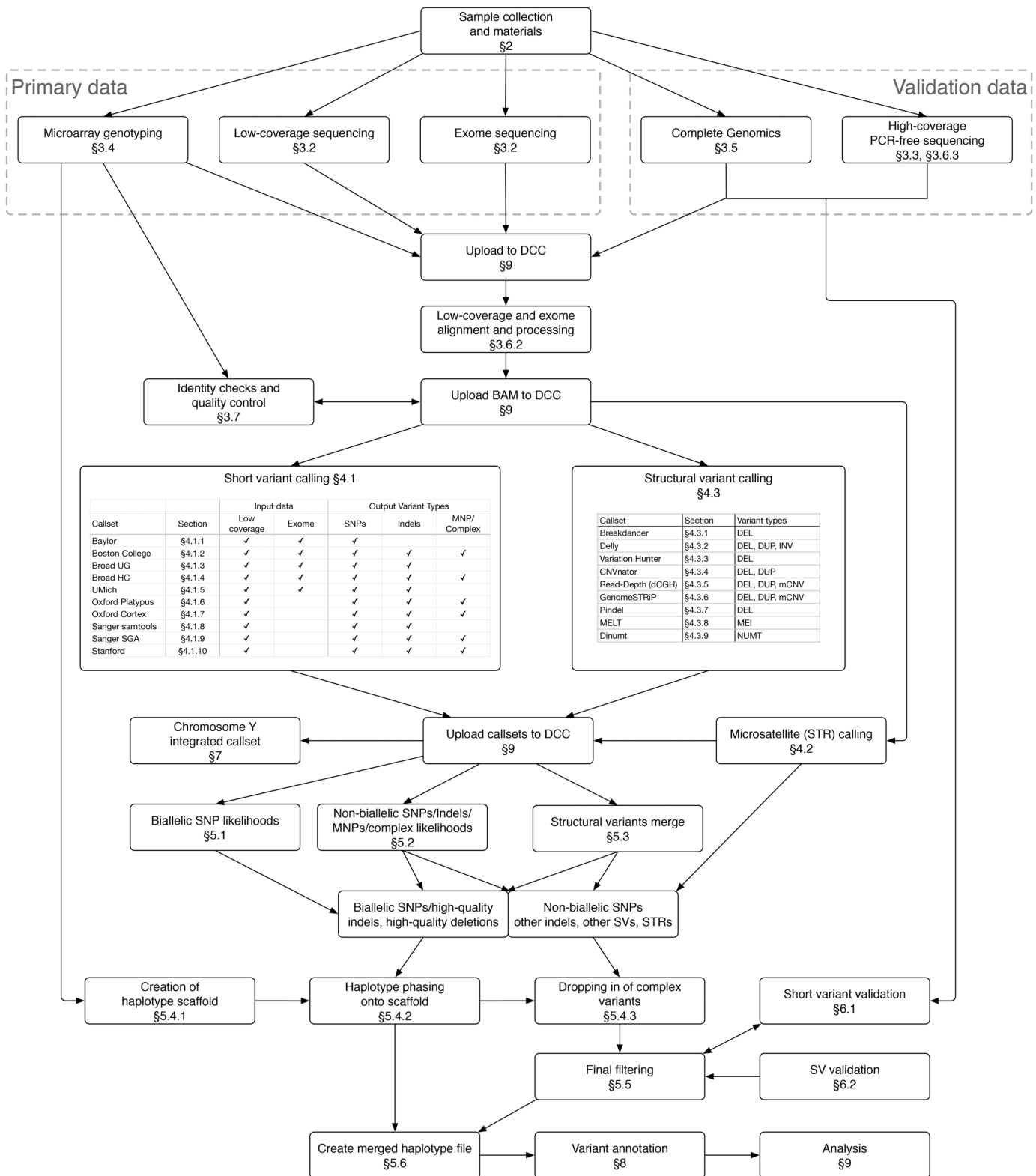
Scientific management: Lisa D. Brooks¹⁶⁶, Adam L. Felsenfeld¹⁶⁶, Jean E. McEwen¹⁶⁶, Yekaterina Vaydylevich¹⁶⁶, Eric D. Green¹⁵, Audrey Duncanson¹⁶⁷, Michael Dunn¹⁶⁷, Jeffery A. Schloss¹⁶⁶, Jun Wang^{26,27,28,29,30}, Huanming Yang^{26,168}

Writing group: Adam Auton¹, Lisa D. Brooks¹⁶⁶, Richard M. Durbin⁴, Erik P. Garrison⁴, Hyun Min Kang², Jan O. Korbel^{12,17}, Jonathan L. Marchini^{8,9}, Shane McCarthy⁴, Gil A. McVean^{8,9}, Gonçalo R. Abecasis²

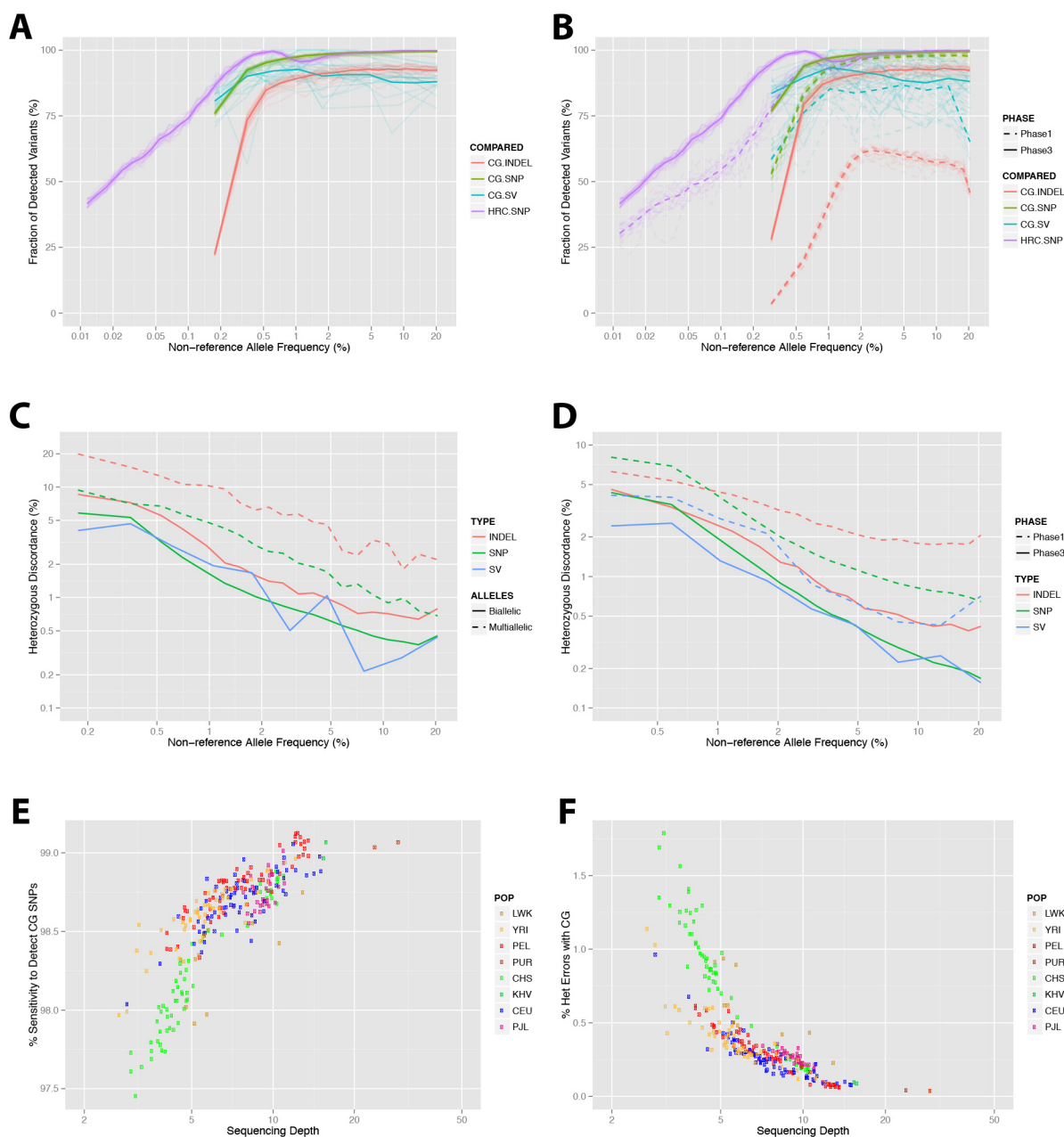
¹Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA. ²Center for Statistical Genetics, Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. ³Vertex Pharmaceuticals, Boston, Massachusetts 02210, USA. ⁴Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK. ⁵Illumina United Kingdom, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK. ⁶McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ⁷Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA. ⁸Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. ⁹Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. ¹⁰Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. ¹¹Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. ¹²European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. ¹³The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ¹⁴Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas 77030, USA. ¹⁵US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA. ¹⁶Centre of Genomics and Policy, McGill University, Montreal, Quebec H3A 1A4, Canada. ¹⁷European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany. ¹⁸The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, Connecticut 06032, USA. ¹⁹Department of Life Sciences, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea 120-750. ²⁰Max Planck Institute for Molecular Genetics, D-14195 Berlin-Dahlem, Germany. ²¹Dahlem Centre for Genome Research and Medical Systems Biology, D-14195 Berlin-Dahlem, Germany. ²²McDonnell Genome Institute at Washington University, Washington University School of Medicine, St Louis, Missouri 63108, USA. ²³USTAR Center for Genetic Discovery & Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA. ²⁴Affymetrix, Santa Clara, California 95051, USA. ²⁵US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA. ²⁶BGI-Shenzhen, Shenzhen 518083, China. ²⁷Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. ²⁸Princess Al Jawhara Albrahman Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 80205, Saudi Arabia. ²⁹Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. ³⁰Department of Medicine and State Key Laboratory for Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong. ³¹Coriell Institute for Medical Research, Camden, New Jersey 08103, USA. ³²European Centre for Public Health Genomics, UNU-MERIT, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands. ³³Alacris Theranostics, D-14195 Berlin-Dahlem, Germany. ³⁴Personalis, Menlo Park, California 94025, USA. ³⁵US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA. ³⁶Department of Computer Engineering, Bilkent University, TR-06800 Bilkent, Ankara, Turkey. ³⁷Seven Bridges Genomics, 1 Broadway, 14th floor, Cambridge, Massachusetts 02142, USA. ³⁸Department of Agronomy, Kansas State University, Manhattan, Kansas 66506, USA. ³⁹Illumina, San Diego, California 92122, USA.

- ⁴⁰Department of Genetics, Harvard Medical School, Cambridge, Massachusetts 02142, USA. ⁴¹SynapDx, Four Hartwell Place, Lexington, Massachusetts 02421, USA. ⁴²Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. ⁴³Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA. ⁴⁴Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA. ⁴⁵Department of Genetic Medicine, Weill Cornell Medical College, New York, New York 10044, USA. ⁴⁶European Molecular Biology Laboratory, Genomics Core Facility, Meyerhofstrasse 1, 69117 Heidelberg, Germany. ⁴⁷Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London WC1E 6DD, UK. ⁴⁸Center for Systems Biology and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ⁴⁹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ⁵⁰Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. ⁵¹Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1498, New York, New York 10029-6574, USA. ⁵²Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA. ⁵³Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁵⁴McGill University and Genome Quebec Innovation Centre, 740, Avenue du Dr. Penfield, Montreal, Quebec H3A 0G1, Canada. ⁵⁵National Eye Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. ⁵⁶New York Genome Center, 101 Avenue of the Americas, 7th floor, New York, New York 10013, USA. ⁵⁷Department of Systems Biology, Columbia University, New York, NY 10032, USA. ⁵⁸Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, New York, USA. ⁵⁹Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139, USA. ⁶⁰General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. ⁶¹Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA. ⁶²Ontario Institute for Cancer Research, MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario, M5G 0A3, Canada. ⁶³Department of Anthropology, Penn State University, University Park, Pennsylvania 16802, USA. ⁶⁴Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey 08903, USA. ⁶⁵Department of Genetics, Stanford University, Stanford, California 94305, USA. ⁶⁶Departments of Genetics and Pathology, Stanford University, Stanford, California 94305-5324, USA. ⁶⁷Ancestry.com, San Francisco, California 94107, USA. ⁶⁸DNAexus, 1975 West El Camino Real STE 101, Mountain View California 94040, USA. ⁶⁹Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO), CINVESTAV, Irapuato, Guanajuato 36821, Mexico. ⁷⁰Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. ⁷¹Department of Microbiology, Tel-Aviv University, Tel-Aviv 69978, Israel. ⁷²International Computer Science Institute, Berkeley, California 94704, USA. ⁷³Thermo Fisher Scientific, 200 Oyster Point Boulevard, South San Francisco, California 94080, USA. ⁷⁴The Translational Genomics Research Institute, Phoenix, Arizona 85004, USA. ⁷⁵Life Technologies, Beverly, Massachusetts 01915, USA. ⁷⁶Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California 90024, USA. ⁷⁷Department of Psychiatry, University of California, San Diego, La Jolla, California 92093, USA. ⁷⁸Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA. ⁷⁹Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA. ⁸⁰Departments of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California 94158, USA. ⁸¹Institute for Quantitative Biosciences (QB3), University of California, San Francisco, 1700 4th Street, San Francisco, California 94158, USA. ⁸²Institute for Human Genetics, University of California, San Francisco, 1700 4th Street, San Francisco, California 94158, USA. ⁸³Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California 95064, USA. ⁸⁴Howard Hughes Medical Institute, Santa Cruz, California 95064, USA. ⁸⁵Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. ⁸⁶Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK. ⁸⁷Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. ⁸⁸Institute for Genetics and Genomics in Geneva, University of Geneva, 1211 Geneva, Switzerland. ⁸⁹Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. ⁹⁰Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. ⁹¹Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁹²Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. ⁹³Department of Pediatrics, University of Pittsburgh, Pittsburgh, Pennsylvania 15224, USA. ⁹⁴The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. ⁹⁵Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. ⁹⁶University of Michigan Sequencing Core, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁹⁷Istituto di Ricerca Genetica e Biomedica, CNR, Monserrato, 09042 Cagliari, Italy. ⁹⁸Dipartimento di Scienze Biomediche, Università degli Studi di Sassari, 07100 Sassari, Italy. ⁹⁹University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, Texas 75390, USA. ¹⁰⁰Department of Pediatrics, University of Montreal, Ste. Justine Hospital Research Centre, Montreal, Quebec H3T 1C5, Canada. ¹⁰¹Department of Genetics, Department of Biostatistics, Department of Computer Science, University of Chapel Hill, North Carolina 27599, USA. ¹⁰²Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, North Carolina 28223, USA. ¹⁰³Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands. ¹⁰⁴Department of Biology, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico 00680, USA. ¹⁰⁵Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA. ¹⁰⁶Department of Genetics, Rutgers University, Piscataway, New Jersey 08854, USA. ¹⁰⁷Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA. ¹⁰⁸Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA. ¹⁰⁹Department of Physiology and Biophysics, Weill Cornell Medical College, New York, New York 10065, USA. ¹¹⁰Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Centre for Neuroscience, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands. ¹¹¹Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University, 6500 HB Nijmegen, The Netherlands. ¹¹²Institute of Genetics and Biophysics, National Research Council (CNR), 80125 Naples, Italy. ¹¹³Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA. ¹¹⁴Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA. ¹¹⁵Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. ¹¹⁶Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, USA. ¹¹⁷Department of Chemistry, Yale University, New Haven, Connecticut 06520, USA. ¹¹⁸Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center 2333 ZA, The Netherlands. ¹¹⁹Department of Computer Science, University of California, San Diego, La Jolla, California 92093, USA. ¹²⁰Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego, La Jolla, California 92093, USA. ¹²¹Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77230, USA. ¹²²Bina Technologies, Roche Sequencing, Redwood City, California 94065, USA. ¹²³Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ¹²⁴Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet Building 208, 2800 Lyngby, Denmark. ¹²⁵Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York 10024, USA. ¹²⁶Department of Invertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA. ¹²⁷School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4701, USA. ¹²⁸Program in Biomedical Informatics, Stanford University, Stanford, California 94305, USA. ¹²⁹Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072, Australia. ¹³⁰Virginia Bioinformatics Institute, 1015 Life Sciences Drive, Blacksburg, Virginia 24061, USA. ¹³¹Division of Allergy and Clinical Immunology, School of Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA. ¹³²Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794, USA. ¹³³Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK. ¹³⁴Genetic Alliance, London N1 3QP, UK. ¹³⁵The Ethox Center, Nuffield Department of Population Health, University of Oxford, Old Road Campus, OX3 7LF, UK. ¹³⁶Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ¹³⁷Department of Medical History and Bioethics, Morgridge Institute for Research, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. ¹³⁸University of Wisconsin Law School, Madison, Wisconsin 53706, USA. ¹³⁹US National Institutes of Health, Center for Research on Genomics and Global Health, National Human Genome Research Institute, 12 South Drive, Bethesda, Maryland 20892, USA. ¹⁴⁰Department of African & African American Studies, Duke University, Durham, North Carolina 27708, USA. ¹⁴¹Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA. ¹⁴²Department of Psychiatry and Clinical Psychobiology & Institute for Brain, Cognition and Behavior (IR3C), University of Barcelona, 08035 Barcelona, Spain. ¹⁴³Cancer and Immunogenetics Laboratory, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK. ¹⁴⁴Laboratory of Molecular Genetics, Institute of Biology, University of Antioquia, Medellín, Colombia. ¹⁴⁵Peking University Shenzhen Hospital, Shenzhen, 518036, China. ¹⁴⁶Institute of Medical Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Kunming 650118, China. ¹⁴⁷Instituto de Biología Molecular y Celular del Cancer, Centro de Investigación del Cancer/IBMCC (CSIC-USAL), Institute of Biomedical Research of Salamanca (IBSAL) & National DNA Bank Carlos III, University of Salamanca, 37007 Salamanca, Spain. ¹⁴⁸Ponce Research Institute, Ponce Health Sciences University, Ponce 00716, Puerto Rico. ¹⁴⁹Chronic Disease Research Centre, Tropical Medicine Research Institute, Cave Hill Campus, The University of the West Indies. ¹⁵⁰Faculty of Medical Sciences, Cave Hill Campus, The University of the West Indies. ¹⁵¹Tropical Metabolism Research Unit, Tropical Medicine Research Institute, Mona Campus, The University of the West Indies. ¹⁵²International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh. ¹⁵³Xishuangbanna Health School, Xishuangbanna 666100, China. ¹⁵⁴Irrua Specialist Teaching Hospital, Edo State, Nigeria. ¹⁵⁵Redeemers University, Ogun State, Nigeria. ¹⁵⁶Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA. ¹⁵⁷Medical Research Council Unit, The Gambia, Atlantic Boulevard, Fajara, P.O. Box 273, Banjul, The Gambia. ¹⁵⁸NHLL, Imperial College London, Hammersmith Hospital, London SW7 2AZ, UK. ¹⁵⁹Centre for Tropical Medicine, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam. ¹⁶⁰Peter Doherty Institute of Infection and Immunity, The University of Melbourne, 792 Elizabeth Street, Melbourne VIC 3000, Australia. ¹⁶¹Kenema Government Hospital, Ministry of Health and Sanitation, Kenema, Sierra Leone. ¹⁶²Tulane University Health Sciences Center, New Orleans, Louisiana 70112, USA. ¹⁶³Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Peru. ¹⁶⁴Center for Non-Communicable Diseases, Karachi, Pakistan. ¹⁶⁵Department of Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ¹⁶⁶US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA. ¹⁶⁷Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. ¹⁶⁸James D. Watson Institute of Genome Sciences, Hangzhou 310008, China.

‡Deceased

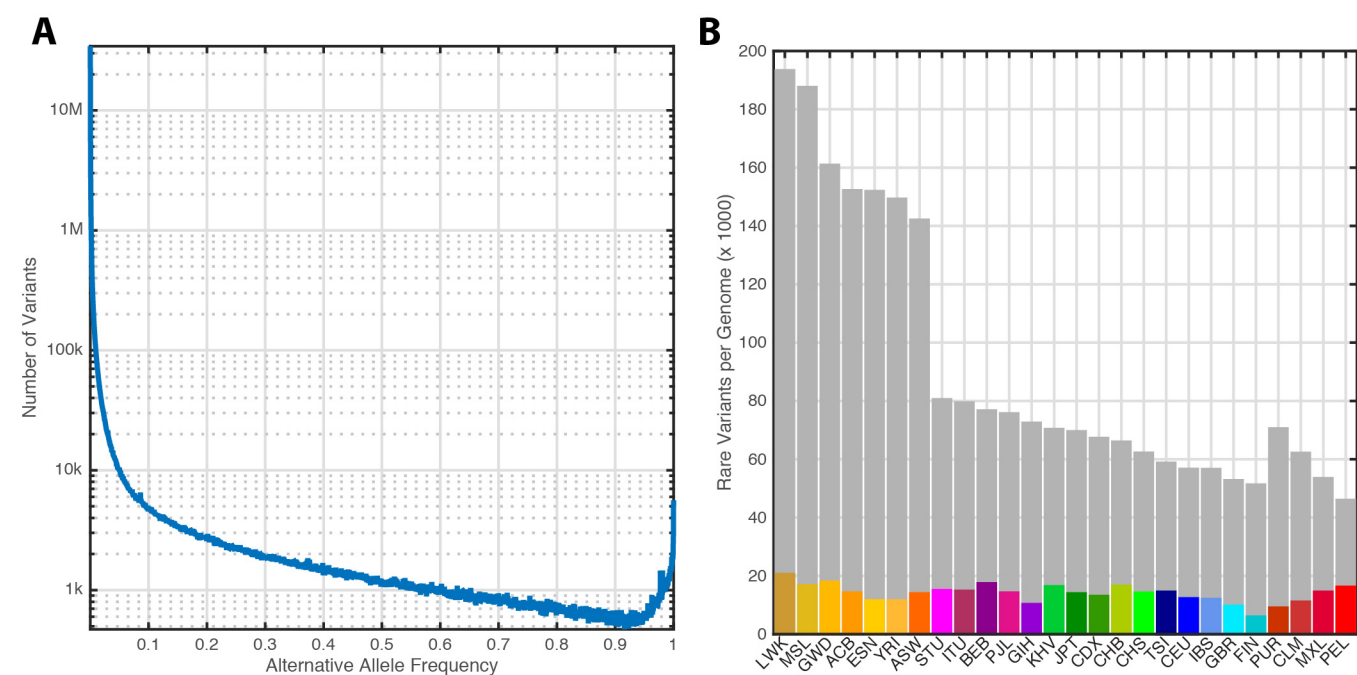


Extended Data Figure 1 | Summary of the callset generation pipeline. Boxes indicate steps in the process and numbers indicate the corresponding section(s) within the Supplementary Information.

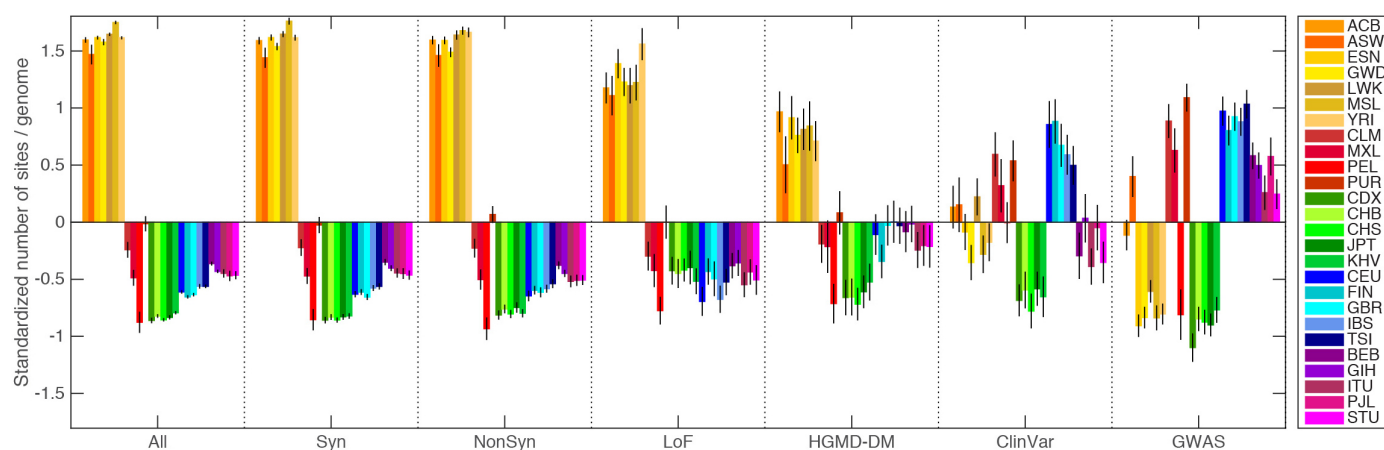


Extended Data Figure 2 | Power of discovery and heterozygote genotype discordance. **a**, The power of discovery within the main data set for SNPs and indels identified within an overlapping sample of 284 genomes sequenced to high coverage by Complete Genomics (CG), and against a panel of >60,000 haplotypes constructed by the Haplotype Reference Consortium (HRC)⁹. To provide a measure of uncertainty, one curve is plotted for each chromosome. **b**, Improved power of discovery in phase 3 compared to phase 1, as assessed in a

sample of 170 Complete Genomics genomes that are included in both phase 1 and phase 3. **c**, Heterozygote discordance in phase 3 for SNPs, indels, and SVs compared to 284 Complete Genomics genomes. **d**, Heterozygote discordance for phase 3 compared to phase 1 within the intersecting sample. **e**, Sensitivity to detect Complete Genomics SNPs as a function of sequencing depth. **f**, Heterozygote genotype discordance as a function of sequencing depth, as compared to Complete Genomics data.

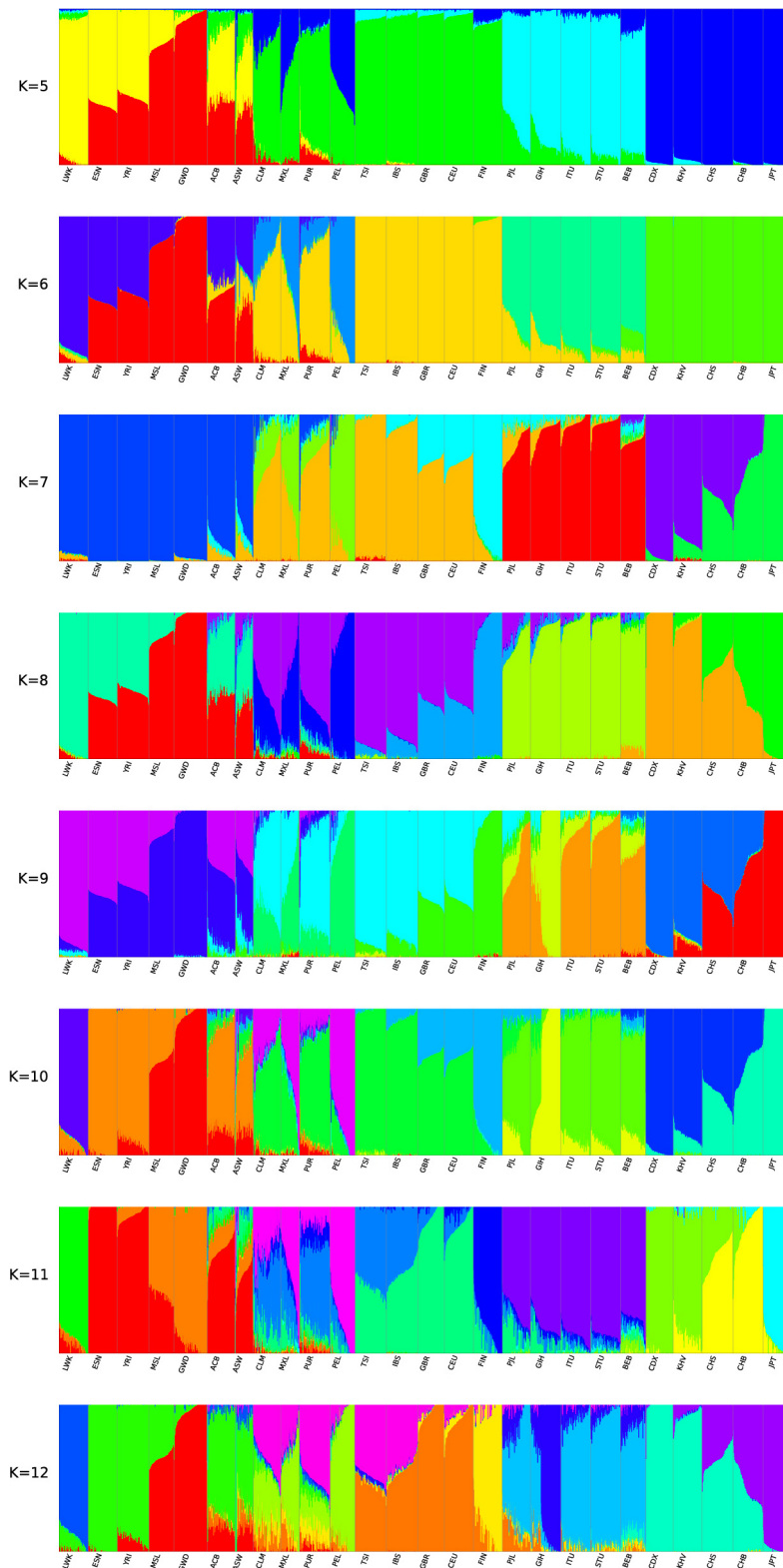


Extended Data Figure 3 | Variant counts. **a**, The number of variants within the phase 3 sample as a function of alternative allele frequency. **b**, The average number of detected variants per genome with whole-sample allele frequencies <0.5% (grey bars), with the average number of singletons indicated by colours.

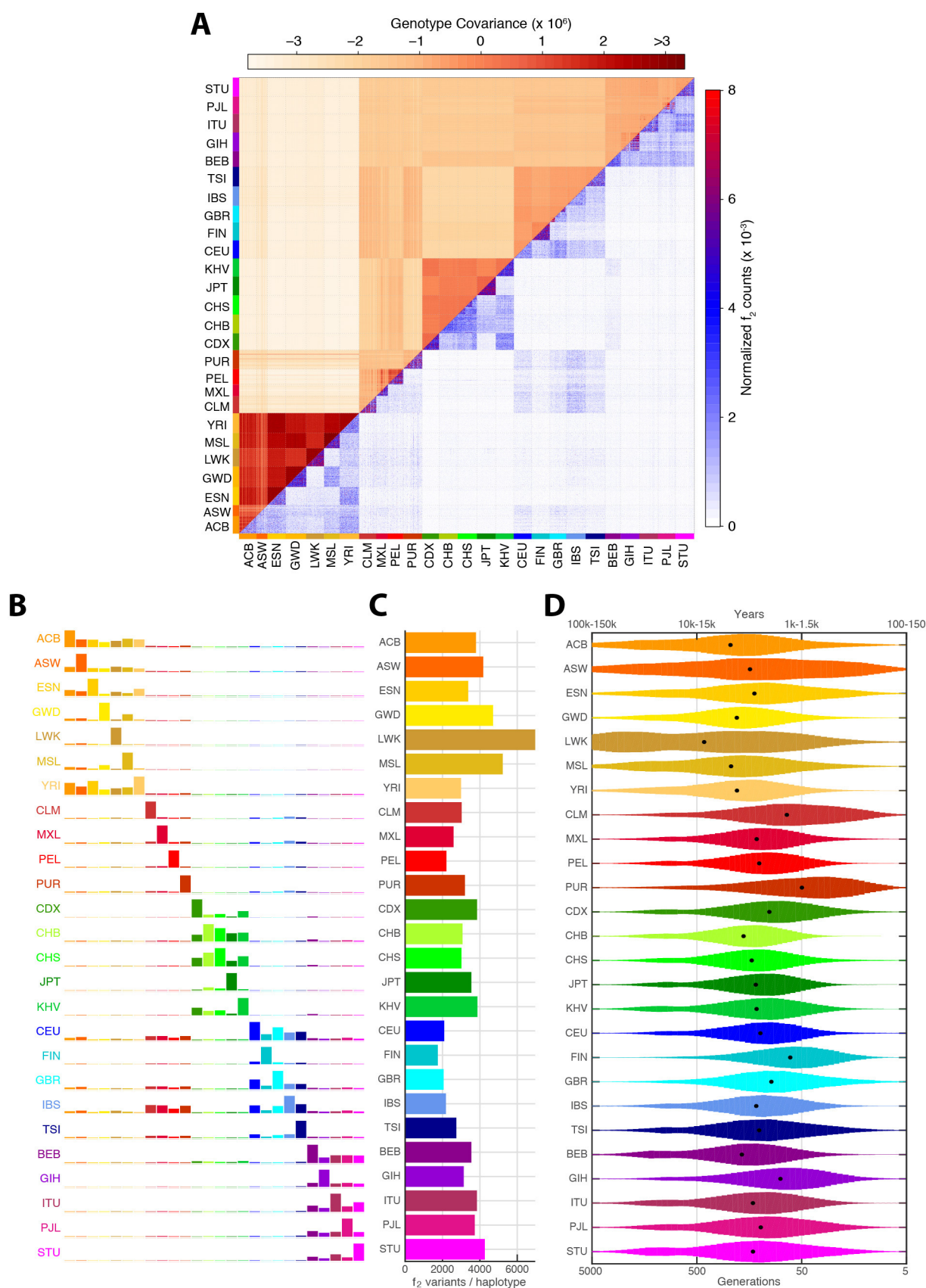


Extended Data Figure 4 | The standardized number of variant sites per genome, partitioned by population and variant category. For each category, z-scores were calculated by subtracting the mean number of sites per genome (calculated across the whole sample), and dividing by the standard deviation.

From left: sites with a derived allele, synonymous sites with a derived allele, nonsynonymous sites with a derived allele, sites with a loss-of-function allele, sites with a HGMD disease mutation allele, sites with a ClinVar pathogenic variant, and sites carrying a GWAS risk allele.

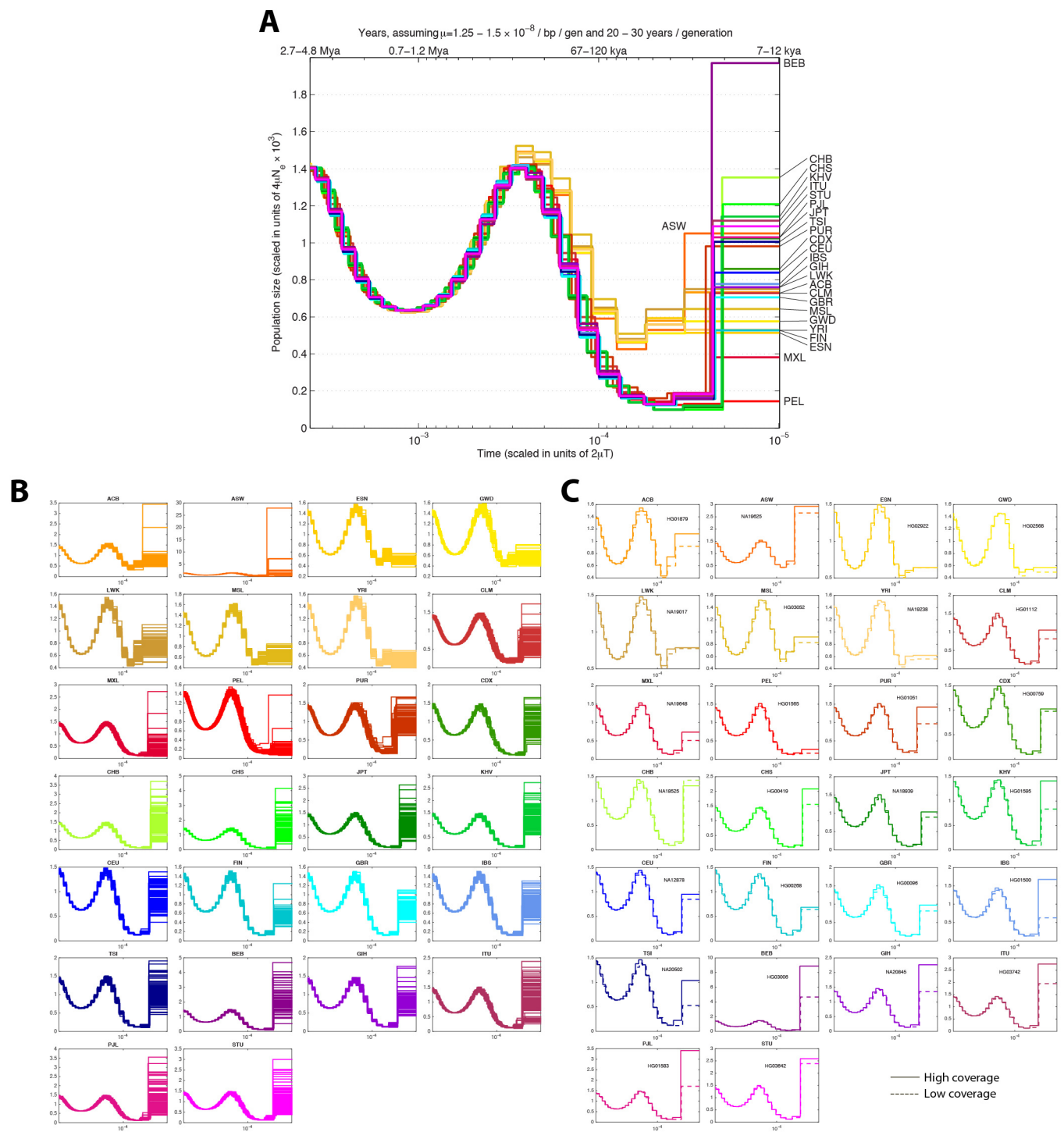


Extended Data Figure 5 | Population structure as inferred using the admixture program for $K = 5$ to 12 .



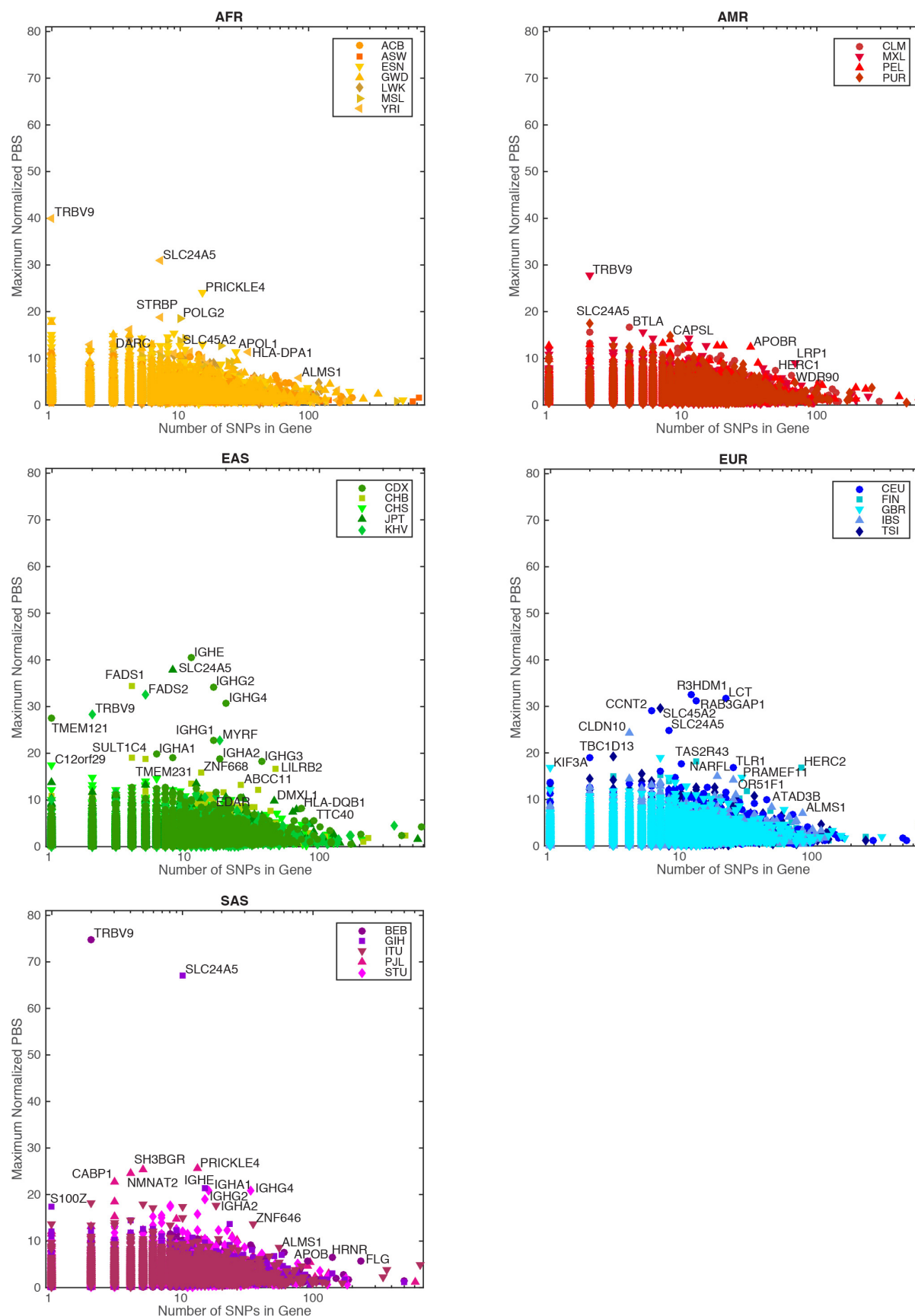
Extended Data Figure 6 | Allelic sharing. **a**, Genotype covariance (above diagonal) and sharing of f_2 variants (below diagonal) between pairs of individuals. **b**, Quantification of average f_2 sharing between populations. Each row represents the distribution of f_2 variants shared between individuals from

the population indicated on the left to individuals from each of the sampled populations. **c**, The average number of f_2 variants per haploid genome. **d**, The inferred age of f_2 variants, as estimated from shared haplotype lengths, with black dots indicating the median value.



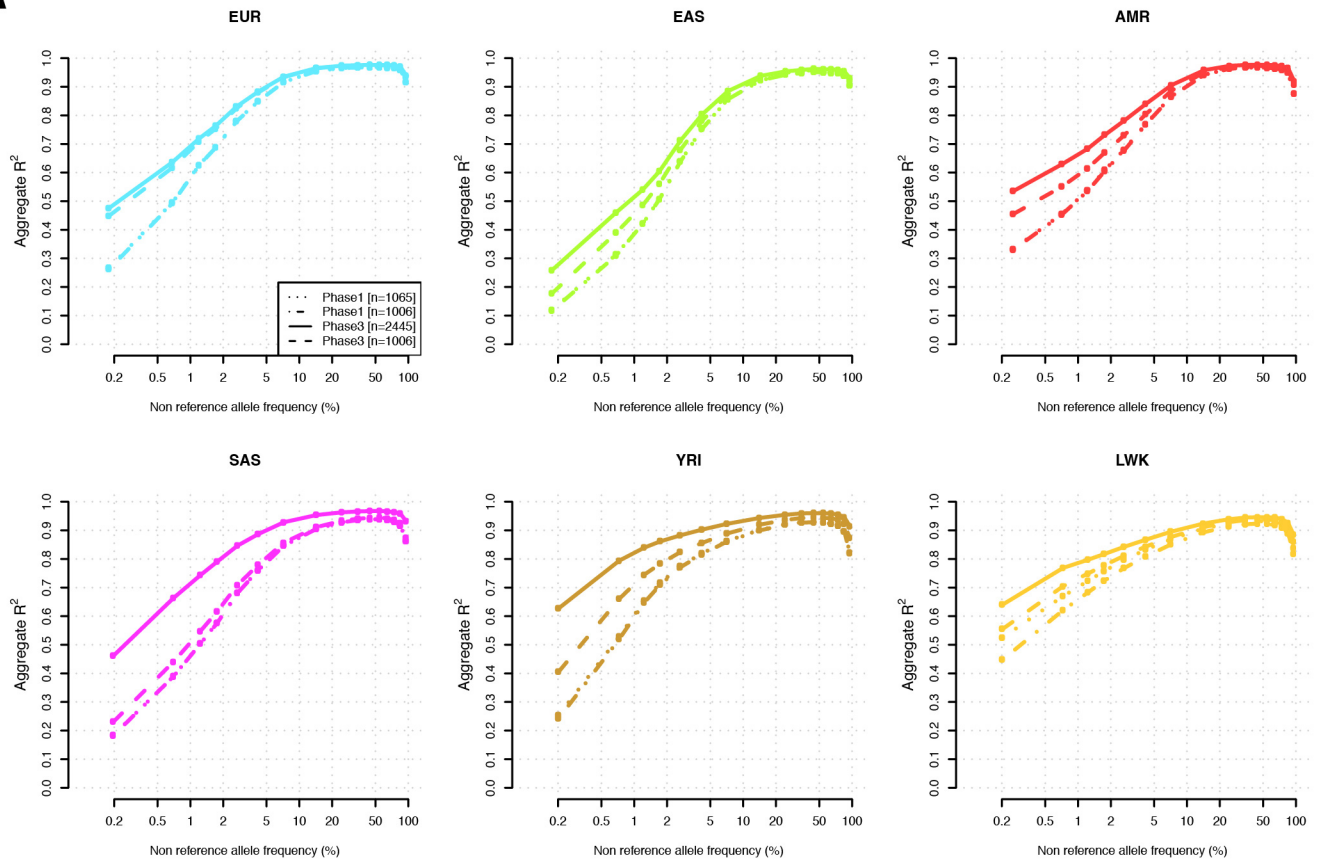
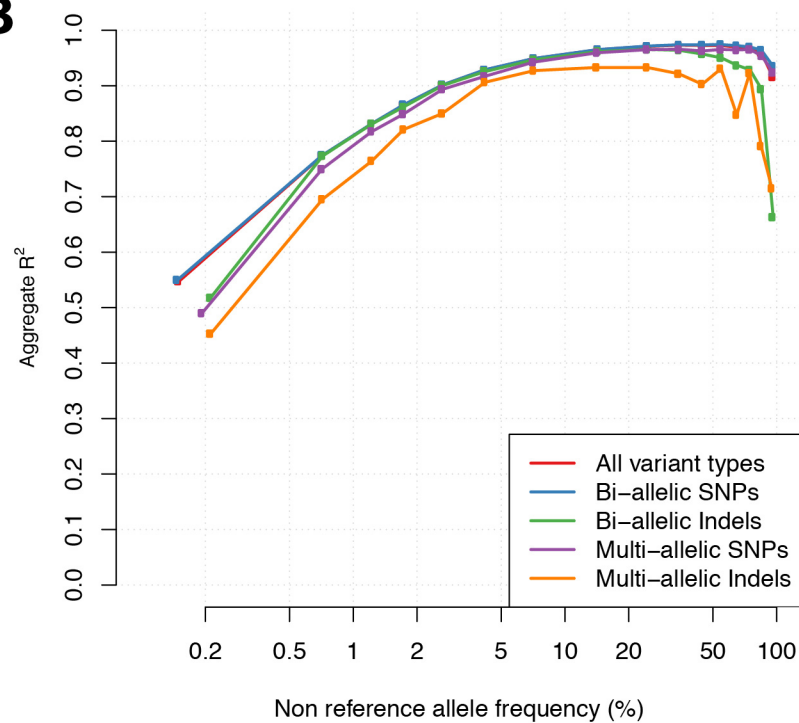
Extended Data Figure 7 | Unsmoothed PSMC curves. **a**, The median PSMC curve for each population. **b**, PSMC curves estimated separately for all individuals within the 1000 Genomes sample. **c**, Unsmoothed PSMC curves comparing estimates from the low coverage data (dashed lines) to those

obtained from high coverage PCR-free data (solid lines). Notable differences are confined to very recent time intervals, where the additional rare variants identified by deep sequencing suggest larger population sizes.



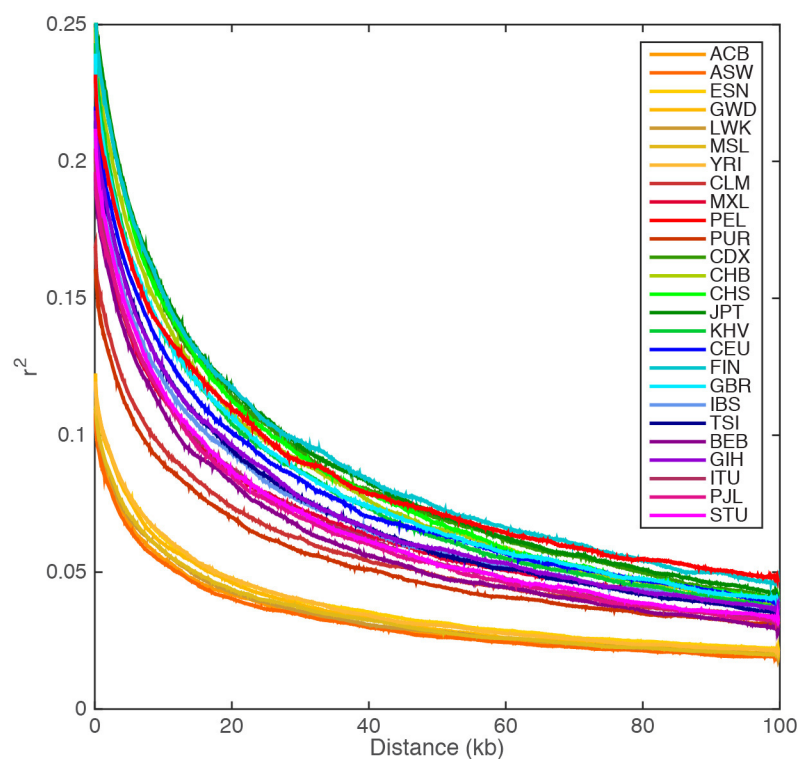
Extended Data Figure 8 | Genes showing very strong patterns of differentiation between pairs of closely related populations within each continental group. Within each continental group, the maximum PBS statistic

was selected from all pairwise population comparisons within the continental group against all possible out-of-continent populations. Note the x axis shows the number of polymorphic sites within the maximal comparison.

A**B**

Extended Data Figure 9 | Performance of imputation. **a**, Performance of imputation in 6 populations using a subset of phase 3 as a reference panel ($n = 2,445$), phase 1 ($n = 1,065$), and the corresponding data within

intersecting samples from both phases ($n = 1,006$). **b**, Performance of imputation from phase 3 by variant class.



Extended Data Figure 10 | Decay of linkage disequilibrium as a function of physical distance. Linkage disequilibrium was calculated around 10,000 randomly selected polymorphic sites in each population, having first thinned

each population down to the same sample size (61 individuals). The plotted line represents a 5 kb moving average.

An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

Structural variants are implicated in numerous diseases and make up the majority of varying nucleotides among human genomes. Here we describe an integrated set of eight structural variant classes comprising both balanced and unbalanced variants, which we constructed using short-read DNA sequencing data and statistically phased onto haplotype blocks in 26 human populations. Analysing this set, we identify numerous gene-intersecting structural variants exhibiting population stratification and describe naturally occurring homozygous gene knockouts that suggest the dispensability of a variety of human genes. We demonstrate that structural variants are enriched on haplotypes identified by genome-wide association studies and exhibit enrichment for expression quantitative trait loci. Additionally, we uncover appreciable levels of structural variant complexity at different scales, including genic loci subject to clusters of repeated rearrangement and complex structural variants with multiple breakpoints likely to have formed through individual mutational events. Our catalogue will enhance future studies into structural variant demography, functional impact and disease association.

Structural variants (SVs), including deletions, insertions, duplications and inversions, account for most varying base pairs (bp) among individual human genomes¹. Numerous studies have implicated SVs in human health with associated phenotypes ranging from cognitive disabilities to predispositions to obesity, cancer and other maladies^{1,2}. Discovery and genotyping of these variants remains challenging, however, since SVs are prone to arise in repetitive regions and internal SV structures can be complex³. This has created challenges for genome-wide association studies (GWAS)^{4,5}. Despite recent methodological and technological advances^{6–9}, efforts to perform discovery, genotyping, and statistical haplotype-block integration of all major SV classes have so far been lacking. Earlier SV surveys depended on microarrays¹⁰ as well as genomic and clone-based approaches limited to a small number of samples^{11–15}. More recently, short-read DNA sequencing data from the initial phases of the 1000 Genomes Project^{8,9} enabled us to construct sets of SVs, genotyped across populations, with enhanced size and breakpoint resolution^{6,7}. Previous 1000 Genomes Project SV set releases, however, encompassed fewer individuals and were largely⁶ or entirely⁸ limited to deletions, in spite of the relevance of other SV classes to human genetics^{1,2,4}.

The objective of the Structural Variation Analysis Group has been to discover and genotype major classes of SVs (defined as DNA variants ≥ 50 bp) in diverse populations and to generate a statistically phased reference panel with these SVs. Here we report an integrated map of 68,818 SVs in unrelated individuals with ancestry from 26 populations (Supplementary Table 1). We constructed this resource by analysing 1000 Genomes Project phase 3 whole-genome sequencing (WGS) data¹⁶ along with data from orthogonal techniques, including long-read single-molecule sequencing (Supplementary Table 2), to characterize hitherto unresolved SV classes. Our study emphasizes the population diversity of SVs, quantifies their functional impact, and highlights previously understudied SV classes, including inversions exhibiting marked sequence complexity.

Construction of our phase 3 SV release

We mapped Illumina WGS data (~ 100 bp reads, mean 7.4-fold coverage) from 2,504 individuals onto an amended version⁸ of the GRCh37 reference assembly using two independent mapping

algorithms—BWA¹⁷ and mrsFAST¹⁸—and performed SV discovery and genotyping using an ensemble of nine different algorithms (Extended Data Fig. 1 and Supplementary Note). We applied several orthogonal experimental platforms for SV set assessment, refinement and characterization (Supplementary Table 2) and to calculate the false discovery rate (FDR) for each SV class (Table 1). Callset refinements facilitated through long-read sequencing enabled us to incorporate a number of additional SVs into our callset, including an additional 698 inversions and 9,132 small (< 1 kbp) deletions, compared to the SV set released with the 1000 Genomes Project marker paper¹⁶. As a result, our callset differs slightly relative to the marker paper's SV set¹⁶ (see Supplementary Table 2). We merged individual callsets to construct our unified release (Table 1), comprising 42,279 biallelic deletions, 6,025 biallelic duplications, 2,929 mCNVs (multi allelic copy-number variants), 786 inversions, 168 nuclear mitochondrial insertions (NUMTs), and 16,631 mobile element insertions (MEIs, including 12,748, 3,048 and 835 insertions of *Alu*, L1 and SVA (SINE-R, VNTR and *Alu* composite) elements, respectively).

SV non-reference genotype concordance estimates ranged from $\sim 98\%$ for biallelic deletions and MEI classes to $\sim 94\%$ for biallelic duplications. 60% of SVs were novel with respect to the Database of Genomic Variants (DGV)¹⁹ (50% reciprocal overlap criterion, Fig. 1a), whereby 71% of SVs (50% reciprocal overlap) and 60% of collapsed copy-number variable regions (CNVRs, 1 bp overlap) were novel compared to previous 1000 Genomes Project releases^{6,8}, reflecting methodological improvements and inclusion of additional populations. Novel SVs showed enrichment for rare sites, which we detected down to an autosomal allele count of '1'. And while variations in FDR estimates were evident with SV size and VAF (variant allele frequency), we consistently estimated the FDR at $\leq 5.4\%$ when stratifying deletions and duplications by size and frequency, including for rare SVs with VAF $< 0.1\%$ (Extended Data Figs 1, 2). A comparison with deep-coverage Complete Genomics (CG) sequencing data indicated an overall sensitivity of 88% for deletions and 65% for duplications, with the false negatives driven largely by the relatively lowered sensitivity for ascertaining small SVs in Illumina sequencing data (Fig. 1b, Extended Data Fig. 3). The average per-individual sensitivity was similar for deletions (89%) and slightly lower for

Table 1 | Phase 3 extended SV release

| SV class | No. sites | Median size of SV sites (bp) | Median kbp per individual | Median alleles per individual | Site FDR | Biallelic site breakpoint precision (bp) | Genotype concordance (non-ref.) | Sensitivity estimates |
|-------------------------|-----------|------------------------------|---------------------------|-------------------------------|------------|--|---------------------------------|-----------------------|
| Deletion (biallelic) | 42,279 | 2,455 | 5,615 | 2,788 | 2%*-4%† | 15 (±50)** 0.7 (±9.5)†† | 98%¶ | 88%¶ |
| Duplication (biallelic) | 6,025 | 35,890 | 518 | 17 | 1%*-4%† | 683 (±1,350)‡‡ | 94%¶ | 65%¶ |
| mCNV | 2,929 | 19,466 | 11,346 | 340 | 1%*-4%† | — | NA | NA |
| Inversion | 786 | 1,697 | 78 | 37 | 17%§ (9%)‡ | 32 (±47) | 96%§ | 32% |
| MEI | 16,631 | 297 | 691 | 1,218 | 4%‡ | 0.95 (±5.93) | 98% | 83#-96%★ |
| NUMT | 168 | 157 | 3 | 5.3 | 10%‡ | 0.25 (±0.43) | 86.1%‡ | NA |

FDR estimates are based on intensity rank-sum testing⁸ using *Affymetrix SNP6 and †Omni 2.5 arrays, ‡PCR, as well as §long-read, ||PCR-free (250 bp-read) and ¶CG sequencing (CG-based estimates used reciprocal overlaps of 50% and 20% for deletions and duplications, respectively). Estimate by comparing MEIs to all #calls or all ★PCR-validated calls from²⁰ (estimates for individual MEI classes are in Supplementary Table 4). NA, no previous data available. Differences in deletion and duplication counts are driven by size-cutoffs and classification of common duplications as mCNVs²⁷. **Ascertained using read-pairs or read-depth. ††Ascertained with split-reads²³. ‡‡Estimated for tandem duplications. |||Estimated for inversions with paired-end support from both breakpoints.

duplications (50%). For MEI classes, estimated sensitivities ranged from 83–96% (Table 1) compared to the 1000 Genomes Project pilot phase where a different MEI detection tool was used²⁰. For inversions, we estimated an overall sensitivity of 32% based on variants with a positive validation status recorded in the InvFEST database²¹, with an increased sensitivity of 67% for inversions <5 kbp in size.

We performed breakpoint assembly using pooled Illumina WGS and Pacific Biosciences (PacBio) sequencing data²², and additionally performed split-read analysis²³ of short reads, to resolve the fine-resolution breakpoint structure of 37,250 SVs (29,954 deletions, 357 tandem duplications, 6,919 MEIs, and 20 inversions; Supplementary Table 3). Breakpoint assemblies showed a mean boundary precision of 0–15 bp for all SV types, with the exception of inversions and duplications for which we achieved mean precision estimates of 32 bp and 683 bp, respectively (Table 1, Fig. 1c).

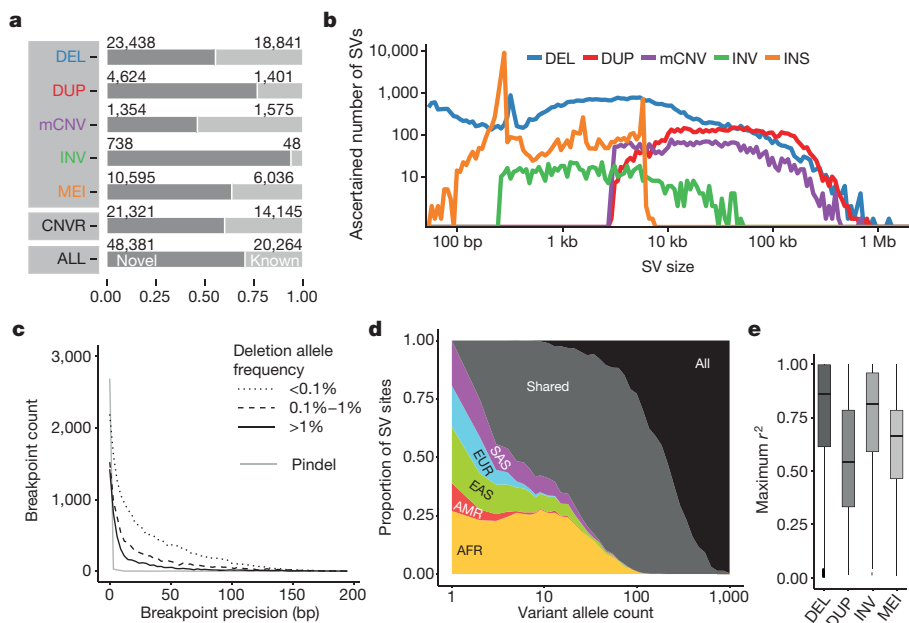
Population genetic properties of SVs

We explored the population genetic properties of SVs among five continental groups—Africa (AFR), the Americas (AMR), East Asia (EAS), Europe (EUR) and South Asia (SAS). The bulk of SVs occur at low frequency (65% exhibit VAF < 0.2%) consistent amongst individual SV classes (Extended Data Figs 2, 3). While rare SVs are typically specific to individual continental groups, at VAF ≥ 2% nearly all SVs are shared across continents (Fig. 1d, Extended Data Fig. 3). Notably, we identified 1,075 SVs with VAF > 50% (889 biallelic deletions, 2 biallelic duplications, 90 mCNVs, 88 MEIs and 6 inversions) encompassing 5 Mbp, sites of interest for future updates to the human reference genome. We estimated the mutation rate for each SV class using Waterson's estimator of θ , for example, ascertaining a mutation rate of 0.113 deletions per haploid genome generation, a threefold higher

estimate compared with previous reports^{10,24}, probably owing to our increased power for detecting variants < 5 kbp (Supplementary Note).

We found that 73% of SVs with >1% VAF and 68% of rarer SVs (VAF > 0.1%) are in linkage disequilibrium (LD) with nearby single nucleotide polymorphisms (SNPs) ($r^2 > 0.6$); however, the proportion of variants in LD highly depends on the SV class (Fig. 1e, Extended Data Fig. 4). For example, only 44% of all biallelic duplications with VAF > 0.1% were in LD with a nearby SNP ($r^2 > 0.6$), in agreement with previous findings^{10,25,26}. Notably, we observed a striking depletion of biallelic duplications amongst common SVs ($P < 2 \times 10^{-16}$, Kolmogorov–Smirnov test; Extended Data Fig. 5) with most common duplications classified as multi-allelic SVs (that is, mCNVs). This behaviour suggests extensive recurrence of SVs at duplication sites consistent with what was recently observed in a smaller cohort of 849 individuals²⁷. These LD characteristics suggest duplications are currently under-ascertained for disease associations using tag-SNP-based approaches.

Based on our haplotype-resolved SV catalogue, we observed that individuals of African ancestry exhibit, on average, 27% more heterozygous deletions than individuals from other populations (mean of 1,705 versus 1,342), consistent with SNPs²⁸ (Extended Data Fig. 5). The relative proportion of deletion- versus SNP-affected sequence, however, showed a 13% excess in non-African compared to African populations (ratio 1.64 versus 1.45). Principal component analyses with different SV classes generally recapitulated continental population structure and admixture (Extended Data Fig. 6 and Supplementary Note). Our analysis further allowed us to identify a catalogue of 6,495 ancestry-informative MEI markers of potential value to population genetics history and forensics research (Extended Data Fig. 5, Supplementary Table 4).



Since population stratification can be used as a signature to detect adaptive selection, we additionally identified SVs varying in VAF amongst different populations. For each SV site we calculated a V_{ST} statistic, a measure highly correlated with F_{ST} (the fixation index)²⁹ that can be applied to assess population stratification of biallelic and multi-allelic SVs²⁹. We observed 1,434 highly stratified SVs ($>0.2 V_{ST}$, corresponding to 2.9 standard deviations (s.d.) from the mean; Supplementary Table 5), among which 578 intersected gene coding sequences (CDSs). Among these were several SVs associated with regions previously reported to be under positive selection, such as *KANSL1* mCNVs (Extended Data Fig. 6) that tag a European-enriched inversion polymorphism associated with increased fecundity³⁰. Most of the population-stratified sites, however, have not been previously described and are, thus, potential targets for future investigation of SVs undergoing adaptive selection or genetic drift. These include, for example, a 14.5 kbp intronic duplication of *HERC2* enriched in East Asians ($V_{ST} = 0.62$ EAS-EUR).

Functional impact of SVs

We analysed the intersection of deletions binned by VAF with various classes of genic and intergenic functional elements (Fig. 2a, Extended Data Fig. 7). The CDSs, untranslated regions (UTRs) and introns of genes, in addition to ENCODE³¹ transcription factor binding sites and ultrasensitive noncoding regions, showed a significant depletion ($P < 0.001$; permutation testing in each VAF bin) compared to a random background model. In general, these elements are more depleted (in terms of fold change) in common VAF bins compared to rarer deletion alleles, in keeping with purifying¹⁰ (or in some cases background³²) selection. Genes more intolerant to mutation (as measured from SNP diversity, residual variation intolerance score

(RVIS)³³ < 20) exhibited the most pronounced depletion ($P < 0.001$; permutation testing between pairs of RVIS-score categories). All other SV classes exhibited similar signatures of selection; when compared to deletions these depletions were, however, more attenuated (Fig. 2b, Extended Data Fig. 7). Additional assessment of the site frequency spectrum showed that, as deletion sizes increase, these SVs become rarer ($P < 2.2 \times 10^{-16}$; linear model, F-test), evidence of purifying selection against events more likely intersecting functional elements. Duplications, by comparison, did not exhibit such trend, consistent with reduced selective constraints (Supplementary Note).

We additionally analysed 5,819 homozygous deletions to search for gene knockouts occurring naturally in human populations. Among these we identified 240 genes (corresponding to 204 individual deletion sites) that, on the basis of the observation of homozygous losses in normal individuals, seem to be 'dispensable' (Supplementary Table 6). Most of the underlying deletions were found in more than one human population, and for only one (0.5%) we observed evidence for the putative involvement of uniparental disomy in the homozygosity (Supplementary Note). The majority ($>80\%$) of these homozygous gene losses were novel compared to a previous analysis based on DGV variants¹⁹, or recent clinical genomics studies (Supplementary Note). As expected, genes affected by homozygous loss were not highly conserved and were relatively tolerant to other forms of genetic variation (RVIS = 0.74 compared to OMIM disease genes showing RVIS = 0.43; $P = 9.4 \times 10^{-25}$; Mann-Whitney test). Moreover, the set was functionally enriched for glycoproteins (Benjamini-Hochberg corrected P -value = 1.6×10^{-3} , EASE (Expression Analysis Systematic Explorer) score) and genes harbouring immunoglobulin domains (Benjamini-Hochberg corrected P -value = 1.0×10^{-5} , EASE score).

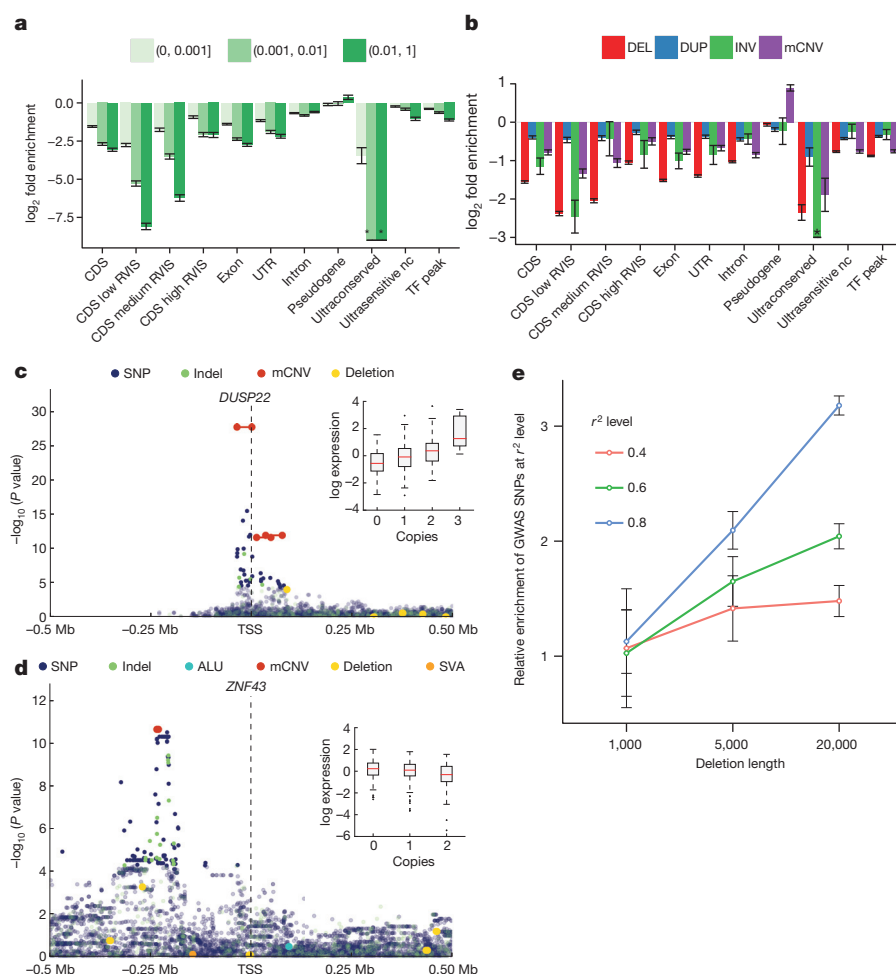


Figure 2 | SV functional impact. **a**, Relative enrichment or depletion of genomic elements within breakpoint-resolved deletions binned by VAF. TF, transcription factor binding site; nc, noncoding. RVIS range from 0–100 (low < 20 , medium 20–50, high ≥ 50). *no element intersected. **b**, Enrichment/depletion of genomic elements within different SV classes, compared with breakpoint-resolved deletions. **c**, Manhattan plot of *DUSP22*-eQTL. Inset, boxplots of association between copy-number genotype and expression. **d**, Manhattan plot of *ZNF43*-eQTL. **e**, Enrichment of SV-containing haplotypes at previously reported GWAS hits (error bars show s.e.m.).

We next quantified the functional impact of SVs using expression quantitative trait loci (eQTL) associations as a surrogate^{34,35}. Based on transcriptome data from lymphoblastoid cell lines derived from 462 individuals³⁶ (the gEUVADIS consortium), we tested 18,969 expressed protein-coding genes for *cis*-eQTL associations, considering 1 Mbp candidate regions upstream and downstream of CDSs. A joint eQTL analysis using SNPs, indels and SVs with VAF >1% identified 54 eQTLs with a lead SV association (denoted SV-eQTL) and 9,537 eQTLs with a lead SNP/indel association (10% FDR). For an additional 166 eQTLs with lead associations to SNPs or indels, we observed SVs in LD ($r^2 > 0.5$) seven times more than when using random variants matched for LD structure, distance to the transcription start site, and VAF, suggesting that a larger number of eQTLs are probably affected by SVs (Extended Data Fig. 8, Supplementary Table 7). In proportion to the number of variants tested, SV classes were up to ~50-fold enriched for SV-eQTLs ($P = 2.84 \times 10^{-39}$, one-sided Fisher's exact test; Supplementary Table 8). Large SVs were associated with increased effect size; for example, a twofold increase in effect size for genic SVs >10 kbp versus variants <1 kbp ($P = 0.0004$; *t*-test; Extended Data Fig. 8). Taken together, although SNPs contribute more eQTLs overall, our results suggest that SVs have a disproportionate impact on gene expression relative to their number.

Among those 220 eQTLs having either an SV-eQTL or an SV in LD with the lead SNP/indel, most were due to deletions (55% of associations), followed by mCNVs (19%) (Supplementary Table 8). Although SV-eQTLs with the largest effect sizes tended to overlap with CDSs, such as for the dual specificity phosphatase 22 (*DUSP22*) gene (Fig. 2c), we also observed several expression-associated SVs strictly intersecting upstream noncoding sequences, including an mCNV upstream of *ZNF43* (Fig. 2d) possibly mediated through variation of a *cis*-regulatory element. We additionally considered the impact of accounting for SVs when constructing personalized reference genomes for transcriptome analysis. To illustrate this, we considered RNA read alignments for the sample NA12878, comparing the standard reference genome with GRCh37-derived personalized references constructed using NA12878 SNPs, or using NA12878 SNPs and SVs. Using such an approach, we observed marked changes in expression for 525 exons (± 10 reads, \geq onefold change relative to the standard reference), 24 of which could be attributed to the inclusion of SVs into the personalized reference (Supplementary Table 9).

The relevance of SVs to eQTLs suggests that a number of disease associations previously detected by GWAS may be attributable to SVs, which are difficult to assess directly in GWAS. To test this hypothesis we compared 12,892 previously reported SNP-based GWAS hits to SVs identified in our data set, identifying 136 candidate SVs in strong LD ($r^2 > 0.8$) with GWAS variants, which represents a 1.5-fold enrichment when compared to a VAF and haplotype size-matched background set and a threefold enrichment for deletions >20 kbp ($P = 0.004$) (Fig. 2e and Supplementary Note). Approximately a third of these candidate GWAS associations (39) were novel, impacting phenotypes such as colorectal cancer and bone mineral density (Supplementary Table 10). Interestingly, 64% of these novel associations were mediated by deletions <1 kbp, a size range for which our study has improved power over previous surveys, which more than doubles (from 18 to 40) the number of SVs <1 kbp in strong LD with a GWAS lead SNP. Thus, our SV resource could facilitate discovery of numerous additional disease-linked SVs.

SV clustering and complexity

Advances in Illumina sequencing towards longer read lengths (~100 bp versus 36 bp)⁶ in conjunction with the population-level data allowed us to perform an in-depth investigation of SV complexity and clustering. We identified 3,163 regions where SVs seemed to cluster (>2 SVs mapping within 500 bp; Supplementary Table 11). To reduce redundancy caused by multiple overlapping calls per sample, we cal-

culated distinct CNVRs per cluster by merging calls per sample and haplotype and then counting the distinct CNVRs produced across samples (average 6.4 ± 7.2 CNVRs per cluster). We identified 30 genomic regions with an excess of CNVRs (>4 s.d. or >36 CNVRs per cluster). This clustering effect was not correlated with segmental duplications ($r = 0.02$) and only partially explained by SNP diversity ($r = 0.15$; Extended Data Fig. 9). CNVR clusters showed enrichment near late-replicating origins ($P = 0.013$, permutation test) and at cytogenetically defined 'fragile' sites ($P = 0.0017$; permutation test). Although the proportion of gene content in regions exhibiting excessive SV clustering was significantly reduced when compared to a null distribution ($P < 0.000001$, permutation test), 1,881 of 3,163 such regions (59%) intersected one or more genes (Supplementary Table 11). This includes a region comprised of 47 SVs (ranking 2nd out of the 30 genomic regions with >4 s.d.) encompassing the pregnancy-specific glycoprotein gene family (Fig. 3a), a set of genes thought to be critically important for maintenance of pregnancy³⁷. Other SV clusters associated with genes (for example, *IMMP2L*, *CHL1* and *GRID2*) have been implicated as potential risk factors for disease, including neurodevelopmental disorders³⁸.

We additionally specifically assessed the complexity of the 29,954 deletions with resolved breakpoints and found that 6% (1,822) intersected another deletion with distinct breakpoints. A larger fraction (16% or 4,813 of assembled deletion sites) showed the presence of additional inserted sequence at deletion breakpoints. We grouped 1,651 deletions with mean size of 3.1 kbp and at least 10 bp of additional DNA sequence between the original SV site boundaries into five broad classes (Fig. 3b, Supplementary Table 12). The most common class ($n = 501$, 30.3%), termed 'Ins with Dup and Del', comprised deletions exhibiting a recognizable duplicated sequence interval within the respective inserted sequence. Notably, in many cases ($n = 191$) the inserted sequences comprised two or more apparent sequence duplications at the deletion boundaries (a class denoted 'Ins with MultiDup and Del'). Additional classes commonly observed include Inv and Del (inversion with adjacent deletion; $n = 9$) and 'MultiDel'—a class where two or more adjacent deletions are separated by at least one sequence 'spacer' of up to ~204 bp in length ($n = 370$). However, not all complex SVs fit into these classes, with 214 sites forming distinct patterns corresponding to multiple classes or exhibiting increased complexity. Template-switching mechanisms could explain the notable complexity of these SVs³. Indeed, microhomology patterns were typically present between the breakpoints of deletions and the respective boundaries of insertion templates at these sites (Extended Data Fig. 9), consistent with formation through single mutational events (Supplementary Note). Across the complex sites assessed, 871 (53%) showed evidence for a local template (≥ 10 bp match, within 10 kbp), whereas for 41 the insertion was presumably templated from a distal region (≥ 22 bp match, >10 kbp away), including 17 sites where the DNA stretch was likely derived from RNA templates (Supplementary Table 13).

To further characterize SV breakpoint complexity, we employed two alternative approaches that do not rely on low-coverage Illumina read assembly. We first examined 7,804 small deletions for breakpoint complexity using split-read analysis²³ (Fig. 3c) and identified 664 (median size 67 bp) exhibiting complexity, 64 of which contained insertions ≥ 3 bp that may be derived from a nearby template (Supplementary Table 14, Extended Data Fig. 9). We additionally realigned long DNA reads from a single individual (NA12878)²² sequenced by high-coverage PacBio (median read length 3.0 kbp) and Molecule (median 3.2 kbp) single-molecule WGS around deletions from our release set (Fig. 3d). Out of 766 deletions in NA12878 investigated with this approach, 62 exhibited complexity showing three to six breakpoints (Supplementary Table 12). A deletion of exon 3 of the serine protease inhibitor *SPINK14*, for example, was accompanied by an inversion of an internal segment of the SV sequence (Fig. 3d, left panel). In contrast to the smaller proportion

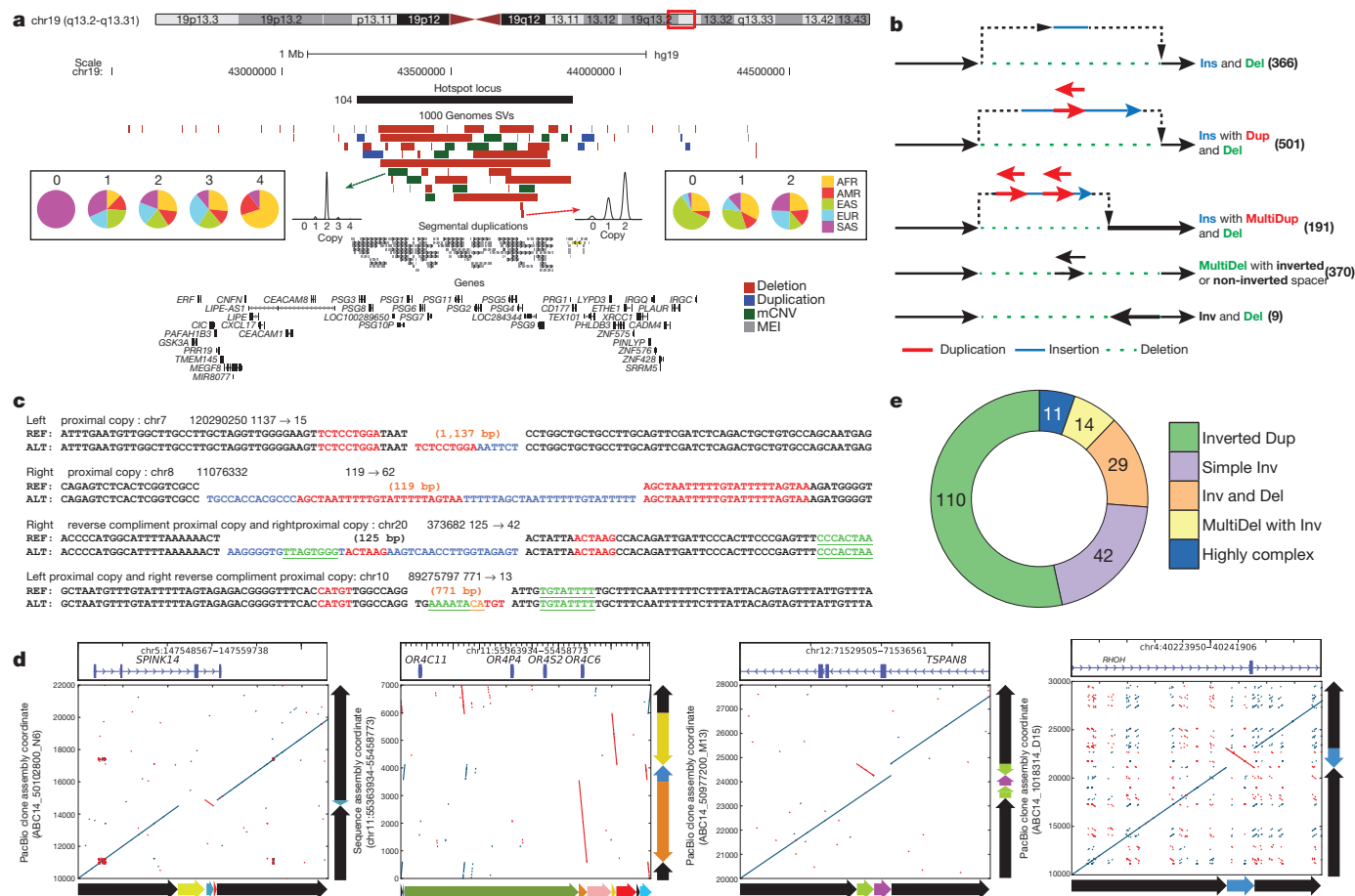


Figure 3 | SV complexity at different scales. **a**, PSG locus with clustered SVs. Population copy-number state histograms are shown for two example SVs. **b**, Schemes depicting assembled complex deletions. **c**, Smaller-scale complex deletions identified with Pindel²³. Flanking sequences are shown for reference (REF) and alternate (ALT) alleles, further to insertions at the breakpoints.

of deletions showing breakpoint complexity, the majority of inversions assessed in NA12878 (19/28) exhibited multiple breakpoints.

To further explore inversion sequence complexity, we performed a battery of targeted analyses, leveraging PacBio resequencing of fosmids (targeting 34 loci), sequencing by Oxford Nanopore Minion (60 loci) and PacBio (206 loci) of long-range PCR amplicons, and data for 13 loci from another sample (CHM1) sequenced by high-coverage PacBio WGS¹⁴. Altogether we verified and further characterized 229 inversion sites, 208 using long-read data and 21 by PCR (Supplementary Table 15), increasing the number of known validated inversions²¹ by >2.5-fold. Remarkably, only 20% of all sequenced inversions characterized in this manner were ‘simple’ (termed ‘Simple Inv’), exhibiting two breakpoints (Fig. 3e), including a 2 kbp inversion on chromosome 4 intersecting a regulatory exon of the Ras homologue family member *RHOH* (Fig. 3d, right panel). The majority of inversions (54%) corresponded to inverted duplications (‘Inverted Dup’; Fig. 3d, middle right panel). In nearly all cases, these involved duplicated stretches <1 kbp inserted within 5 kbp of the alternate copy, suggesting a common mechanism of SV formation (Extended Data Fig. 10). The remaining inversions comprised ‘Inv and Del’ events (14%), ‘MultiDel’ events exhibiting inverted spacers (7%), and more highly complex sites (5%; Fig. 3d, middle left panel). The appreciable inversion complexity uncovered here is most likely due to a mutational process forming complex SVs, potentially involving DNA replication errors³, rather than due to recurrent rearrangement, as our analyses failed to detect corresponding intermediate events in 1000 Genomes Project samples.

Proximal stretches matching the insertion are labelled in red (forward) and green (reverse complement). Blue, insertions lacking nearby matches. **d**, Alignment dot plots depicting inversions (inverted sequences are in red within each dot plot). Adjacent schemes depict allelic structures for REF and ALT. **e**, Inversion complexity summarized.

Discussion

We present what is to our knowledge the most comprehensive set of human SVs to date as an integrated resource for future disease and population genetics studies. We estimate that individuals harbour a median of 18.4 Mbp of SVs per diploid genome, an excess contributed to a large extent by mCNVs (11.3 Mbp) and biallelic deletions (5.6 Mbp; Table 1). When collapsing mCNV sites carrying multiple copies as well as homozygous SVs onto the haploid reference assembly, a median of 8.9 Mbp of sequence are affected by SVs, compared to 3.6 Mbp for SNPs. Furthermore, 37,250 SVs have mapped breakpoints amounting to >113 Mbp of SV sequence resolved at the nucleotide-level. By mining homozygous deletions we identified over two hundred nonessential human genes, a set enriched for immunoglobulin domains that hence may reflect variation in the immune repertoire underlying inter-individual differences in disease susceptibility.

We demonstrate that SV classes are disproportionately enriched (by up to ~50-fold) for SV-eQTLs, although only 220 SVs were found either as lead eQTL association or in high LD with the respective lead SNP. While this corresponds to proportionally fewer associations relative to SNPs compared to a prior estimate based on array technology³⁴, this may be explained by the reliance of this prior estimate on bacterial artificial chromosome arrays, which ascertain large SVs (>50 kbp) that associate with strong effect size, as well as by the relative scarcity of SNPs tested in an earlier study³⁴ (HapMap Phase I)³⁹. We further expand the number of candidate SVs in strong LD with GWAS hits by ~30% (39/136 novel associations implicating SVs as candidates) and find that GWAS haplotypes are enriched up to

threefold for common SVs, which emphasizes the relevance of ascertaining SVs in disease studies. The large number of novel SVs smaller than 1 kbp in length associated with previously reported GWAS hits highlights the importance of increasing sensitivity for SV detection and genotyping at this size range. Additionally, the large number of rare SVs captured by our resource may be of value for disease association studies investigating rare variants.

Our deep population survey has identified hotspots of SV mutation that cannot be accounted for by deep coalescence or segmental duplication content. We describe hitherto undescribed patterns of SV complexity, particularly for inversions. These patterns indicate that other more complex mutational processes outside of non-allelic homologous recombination, retrotransposition, and non-homologous end-joining played an important role in shaping our genome. In spite of this, it remains difficult to fully disentangle the contributions of SV mutation rates and selective forces to the observed variant clustering. The findings presented here leveraged substantial recent technological advances, including increases in Illumina read length and developments in long-read DNA technologies. SV discovery remains a challenge nonetheless, and the full complexity and spectrum of SV is not yet understood. Our analyses, for example, are largely based on 7.4-fold Illumina WGS and, thus, are underpowered to capture much of the complexity of variation, including SVs in repetitive regions, non-reference insertions, and short SVs at the boundaries of the detection limits of read-depth and paired-end-based SV discovery⁴. Furthermore, while many SVs in our callset are statistically phased, the diploid nature of the genome is non-optimally captured by current analysis approaches, which mostly rely on mapping to a haploid reference. We envision that in the future, the use of technology allowing substantial increases in read lengths over the current state-of-the-art will enable genomic analyses of truly diploid sequences to facilitate targeting these additional layers of genomic complexity. Until this is realized, our SV set represents an invaluable resource for the construction and analysis of personalized genomes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 February; accepted 20 August 2015.

- Weischenfeldt, J., Symmons, O., Spitz, F. & Korb, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Rev. Genet.* **14**, 125–138 (2013).
- Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
- Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Rev. Genet.* **10**, 551–564 (2009).
- Alkan, C., Coe, B. P. & Eichler, E. E. Genomic structural variation discovery and genotyping. *Nature Rev. Genet.* **12**, 363–376 (2011).
- Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).
- Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
- Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
- The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
- Kidd, J. M. *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837–847 (2010).
- Korb, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
- Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11**, R52 (2010).
- Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
- Teague, B. *et al.* High-resolution human genome structure by single-molecule analysis. *Proc. Natl Acad. Sci. USA* **107**, 10848–10853 (2010).
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* <http://dx.doi.org/10.1038/nature15393> (this issue).
- Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
- Hach, F. *et al.* mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res.* **42**, W494–W500 (2014).
- MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
- Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* **7**, e1002236 (2011).
- Martinez-Fundichely, A. *et al.* InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res.* **42**, D1027–D1032 (2014).
- Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods* **12**, 780–786 (2015).
- Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
- Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015).
- McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40**, 1166–1174 (2008).
- Locke, D. P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
- Handmaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nature Genet.* **47**, 296–303 (2015).
- Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genet.* **46**, 220–224 (2014).
- Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
- Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
- Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomics. *PLoS Genet.* **9**, e1003709 (2013).
- Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
- Schlattl, A., Anders, S., Waszak, S. M., Huber, W. & Korb, J. O. Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res.* **21**, 2004–2013 (2011).
- Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
- Moore, T. & Dveksler, G. S. Pregnancy-specific glycoproteins: complex gene families regulating maternal-fetal interactions. *Int. J. Dev. Biol.* **58**, 273–280 (2014).
- Girirajan, S. *et al.* Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet.* **7**, e1002334 (2011).
- International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
- Conrad, D. F. *et al.* Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nature Genet.* **42**, 385–391 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Hurles, R. Durbin and D. Reich for valuable comments during the preparation of this work, S. Scherer for providing PCR-based inversion genotyping data for the initial calibration of our inversion caller, B. Nelson and V. Benes for technical assistance, and T. Brown and N. Habermann for critical review of the manuscript. The following people are acknowledged for contributing to PacBio sequencing or analysis: E. Patel, S. Lee, H. Doddapaneni, L. Lewis, R. Ruth, Q. Meng, V. Vee, Y. Han, J. Jayaseelan, A. English, J. Korlach, M. Hunkapiller, B. Hüttel and R. Reinhardt. We acknowledge the Yale University Biomedical High-Performance Computing Center and high-performance compute infrastructure made available through the EMBL and EMBL-EBI IT facilities. We thank the people generously contributing samples to the 1000 Genomes Project. Funding for this research project came from the following grants: NIH U41HG007497 (to C.L., E.E.E., J.O.K., M.A.B., M.G., S.A.M., R.E.M. and J.S.), RO1GM59290 (M.A.B.), RO1HG002898 (S.E.D.) and RO1CA166661 (R.E.D.), PO1HG007497 (to E.E.E.), RO1HG007068 (to R.E.M.), RR19895 and SRO29676-01 (to M.B.G.), Wellcome Trust WT085532/Z/08/Z and WT104947/Z/14/Z (to P.F.), an Emmy Noether Grant from the German Research Foundation (KO4037/1-1, to J.O.K.) and the European Molecular Biology Laboratory. C.L. is on the scientific advisory board (SAB) of BioNano Genomics. E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. and is a consultant for Kunming University of Science and Technology (KUST) as part of the 1000 China Talent Program. P.F. is on the SAB of Omicia, Inc. C.L. is an Ewha Womans University Distinguished Professor. E.E.E. is an investigator of the Howard Hughes Medical Institute. J.O.K. is a European Research Council investigator.

Author Contributions SV discovery & genotyping: R.E.H., P.H.S., T.R., E.J.G., A.A.B., K.Y., F.H., K.C., G.D., K.W., M.H.-Y.F., S.K., C.A., S.A.M., R.E.M., K.Y., M.B.G., S.E.D., E.E.E., J.O.K.; SV merging & haplotype integration: T.R., R.E.H., M.H.-Y.F., E.G., A.Me., S.McC.; SV validation: R.E.H., A.A.B., G.J., M.H.-Y.F., A.M.S., M.K.K., A.Ma., S.K., M.M., M.J.P.C., S.M., P.C., S.E., J.M.K., B.R., J.A.W., F.Y., T.Z., M.A.B., R.E.M., A.B., C.L., E.E.E., J.O.K.; additional analyses: A.Au.,

C.E.M., E.C., E.D., E.-W.L., F.K., J.H., Y.Z., X.S., F.P.C., M.M., M.J.P.C., G.M., S.M., D.A., T.B., J.C., Z.C., L.D., X.F., M.G., J.M.K., H.Y.K.L., Y.K., X.J.M., B.J.N., A.N., R.A.G., M.P., M.R., R.S., D.M.M., M.W., N.F.P., A.Q., E.E.S., A.S., A.A.S., A.U., C.Z., J.Z., W.Z., J.S., O.S.; data management & archiving: L.C., X.Z.-B., P.F.; display items: P.H.S., T.R., E.J.G., A.A., Y.Z., J.H., M.H.-Y.F., K.Y., M.B.G., A.B., O.S., R.E.M., S.E.D., E.E.E., J.O.K.; organization of Supplementary Material: G.D., J.O.K., P.H.S., R.E.M.; SV Analysis group co-chairs: C.L., E.E.E., J.O.K.; manuscript writing: P.H.S., T.R., E.J.G., J.H., R.E.M., M.B.G., O.S., S.E.D., E.E.E., J.O.K.

Author Information Sequencing data, archive accessions and supporting datasets including GRCh37 variant call files comprising the extended SV Analysis Group release set, a 'readme' describing differences to the phase 3 marker paper variant release¹⁶, and a GRCh38 version of our callset, are available at <http://www.1000genomes.org/phase-3-structural-variant-dataset>. DGV archive accession: esd219. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.E.E. (eee@gs.washington.edu) or J.O.K. (korbel@embl.de).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

Peter H. Sudmant^{1*}, Tobias Rausch^{2*}, Eugene J. Gardner^{3*}, Robert E. Handsaker^{4,5*}, Alexey Abyzov^{6*}, John Huddleston^{1,7*}, Yan Zhang^{8,9*}, Kai Ye^{10,11*}, Goo Jun^{12,13}, Markus Hsi-Yang Fritz², Miriam K. Konkel¹⁴, Ankit Malhotra¹⁵, Adrian M. Stütz², Xinghua Shi¹⁶, Francesco Paolo Casale¹⁷, Jieming Chen^{8,18}, Fereydoon Hormozdizari¹, Gargi Dayama¹⁹, Ken Chen²⁰, Maika Malig¹, Mark J. P. Chaisson¹, Klaudia Walter²¹, Sascha Meiers², Seva Kashin^{4,5}, Erik Garrison²², Adam Auton²³, Hugo Y. K. Lam²⁴, Xinmeng Jasmine Mu^{8,25}, Can Alkan²⁶, Danny Antaki²⁷, Taejeong Bae⁶, Eliza Cerveira¹⁵, Peter Chines²⁸, Zechen Chong²⁰, Laura Clarke¹⁷, Elif Dal²⁶, Li Ding^{10,11,29,30}, Sarah Emery³¹, Xian Fan²⁰, Madhusudan Gujral²⁷, Fatma Kahveci²⁶, Jeffrey M. Kidd^{12,31}, Yu Kong²³, Eric Wubbo Lameijer³², Shane McCarthy²¹, Paul Flicek¹⁷, Richard A. Gibbs³³, Gabor Marth²², Christopher E. Mason^{34,35}, Androniki Menelaou^{36,37}, Donna M. Muzny³⁸, Bradley J. Nelson¹, Amina Noor²⁷, Nicholas F. Parrish³⁹, Matthew Pendleton³⁸, Andrew Quitadamo¹⁶, Benjamin Raeder², Eric E. Schadt³⁸, Mallory Romanovitch¹⁵, Andreas Schlattl², Robert Sebra³⁸, Andrey A. Shabalov⁴⁰, Andreas Untergasser^{2,41}, Jerilyn A. Walker¹⁴, Min Wang³³, Fuli Yu³³, Chengsheng Zhang¹⁵, Jing Zhang^{8,9}, Xiangqun Zheng-Bradley¹⁷, Wanding Zhou²⁰, Thomas Zichner², Jonathan Sebat²⁷, Mark A. Batzer¹², Steven A. McCarroll^{4,5}, The 1000 Genomes Project Consortium[†], Ryan E. Mills^{19,31}, Mark B. Gerstein^{8,9,42}, Ali Bashir³⁸, Oliver Stegle¹⁷, Scott E. Devine³, Charles Lee^{15,43}, Evan E. Eichler^{1,7} & Jan O. Korbel^{2,17}§

¹Department of Genome Sciences, University of Washington, 3720 15th Avenue NE, Seattle, Washington 98195-5065, USA. ²European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany. ³Institute for Genome Sciences, University of Maryland School of Medicine, 801 W Baltimore Street, Baltimore, Maryland 21201, USA. ⁴Department of Genetics, Harvard Medical School, Boston, 25 Shattuck Street, Boston, Massachusetts 02115, USA. ⁵Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, Massachusetts 02142, USA. ⁶Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, 200 First Street SW, Rochester, Minnesota 55905,

USA. ⁷Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. ⁸Program in Computational Biology and Bioinformatics, Yale University, BASS 432 & 437, 266 Whitney Avenue, New Haven, Connecticut 06520, USA.

⁹Department of Molecular Biophysics and Biochemistry, School of Medicine, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. ¹⁰The Genome Institute, Washington University School of Medicine, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. ¹¹Department of Genetics, Washington University in St Louis, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. ¹²Department of Biostatistics and Center for Statistical Genetics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109, USA. ¹³Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Pressler St., Houston, Texas 77030, USA. ¹⁴Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, Louisiana 70803, USA. ¹⁵The Jackson Laboratory for Genomic Medicine, 10 Discovery 263 Farmington Avenue, Farmington, Connecticut 06030, USA. ¹⁶Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, North Carolina 28223, USA.

¹⁷European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ¹⁸Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, Connecticut 06520, USA. ¹⁹Department of Computational Medicine & Bioinformatics, University of Michigan, 500 S. State Street, Ann Arbor, Michigan 48109, USA. ²⁰The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, USA. ²¹The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ²²Department of Biology, Boston College, 355 Higgins Hall, 140 Commonwealth Avenue, Chestnut Hill, Massachusetts 02467, USA. ²³Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, New York 10461, USA. ²⁴Bina Technologies, Roche Sequencing, 555 Twin Dolphin Drive, Redwood City, California 94065, USA. ²⁵Cancer Program, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, Massachusetts 02142, USA. ²⁶Department of Computer Engineering, Bilkent University, 06800 Ankara, Turkey. ²⁷University of California San Diego (UCSD), 9500 Gilman Drive, La Jolla, California 92093, USA.

²⁸National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892 USA. ²⁹Department of Medicine, Washington University in St Louis, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. ³⁰Siteman Cancer Center, 660 South Euclid Avenue, St Louis, Missouri 63110, USA. ³¹Department of Human Genetics, University of Michigan, 1241 Catherine Street, Ann Arbor, Michigan 48109, USA.

³²Molecular Epidemiology, Leiden University Medical Center, Leiden 2300RA, The Netherlands. ³³Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.

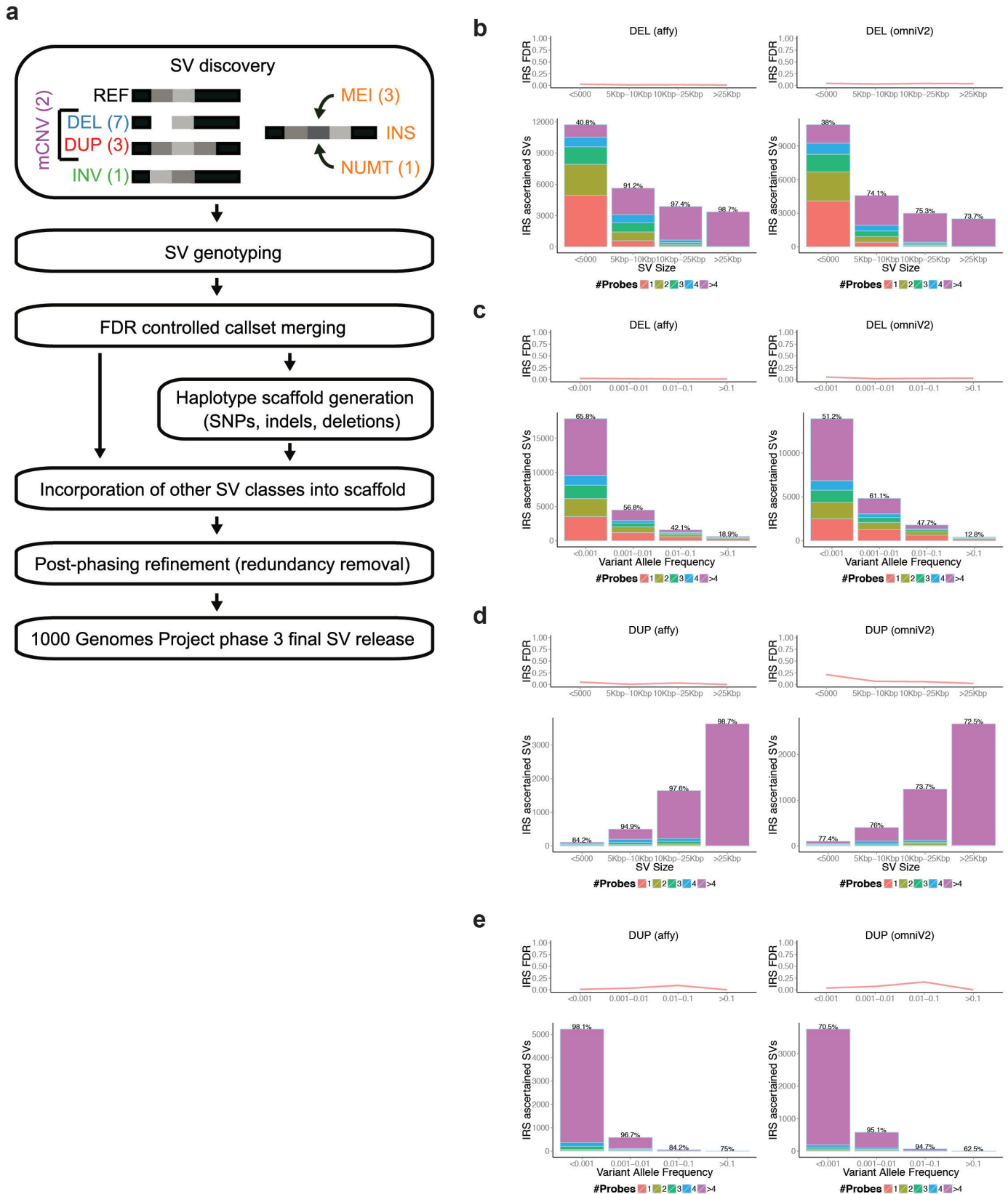
³⁴The Department of Physiology and Biophysics and the HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, 1305 York Avenue, Weill Cornell Medical College, New York, New York 10065, USA. ³⁵The Feil Family Brain and Mind Research Institute, 413 East 69th St, Weill Cornell Medical College, New York, New York 10065, USA. ³⁶University of Oxford, 1 South Parks Road, Oxford OX3 9DS, UK.

³⁷Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, 3584 CG, The Netherlands. ³⁸Department of Genetics and Genomic Sciences, Icahn School of Medicine, New York School of Natural Sciences, 1428 Madison Avenue, New York, New York 10029, USA. ³⁹Institute for Virus Research, Kyoto University, 53 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan. ⁴⁰Center for Biomarker Research and Precision Medicine, Virginia Commonwealth University, 1112 East Clay Street, McGuire Hall, Richmond, Virginia 23298-0581, USA. ⁴¹Zentrum für Molekulare Biologie, University of Heidelberg, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany. ⁴²Department of Computer Science, Yale University, 51 Prospect Street, New Haven, Connecticut 06511, USA. ⁴³Department of Graduate Studies – Life Sciences, Ewha Womans University, Ewhayodae-gil, Seodaemun-gu, Seoul 120-750, South Korea.

*These authors contributed equally to this work.

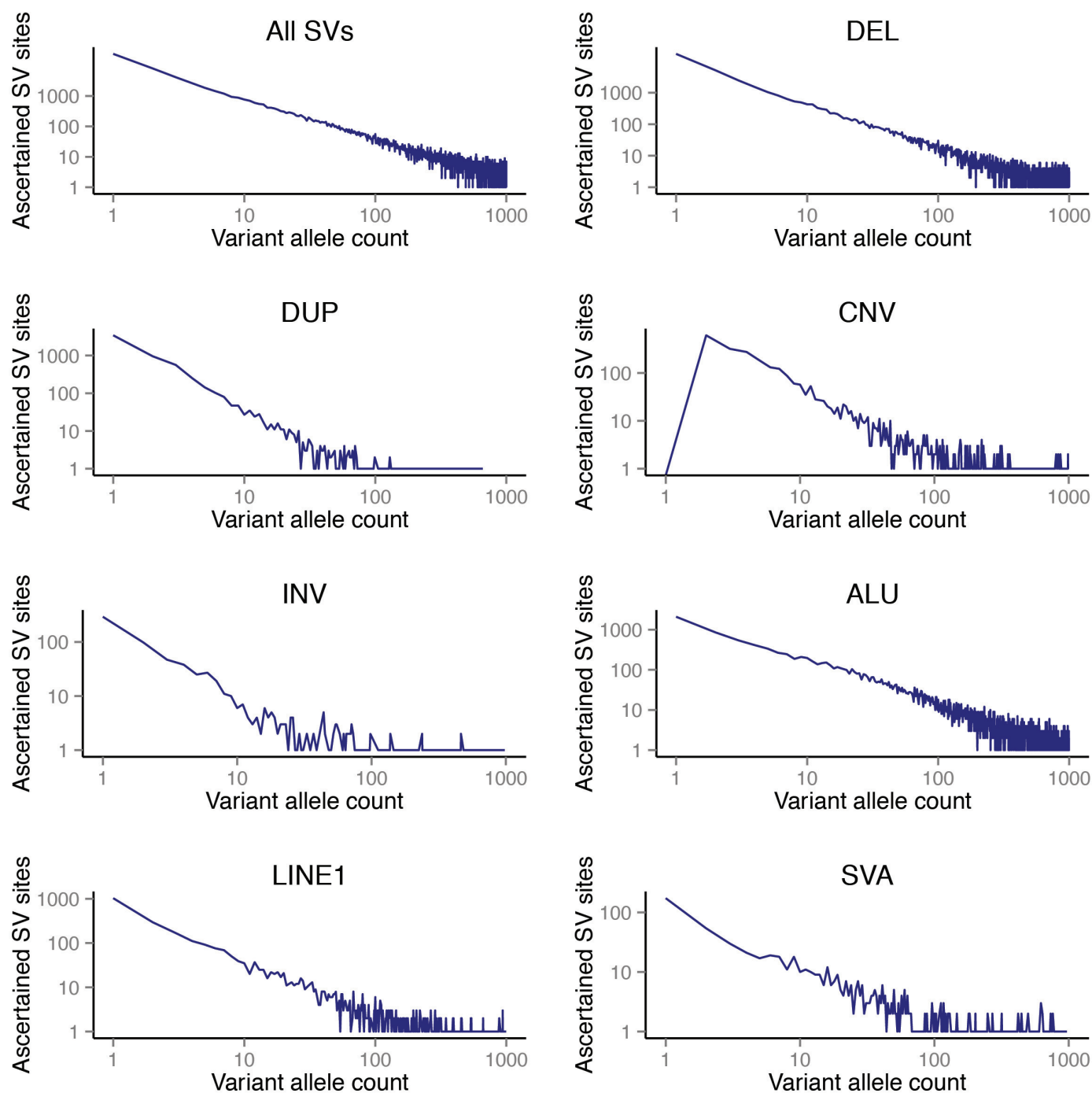
†A list of participants and their affiliations appears in the Supplementary Information.

§These authors jointly supervised this work.



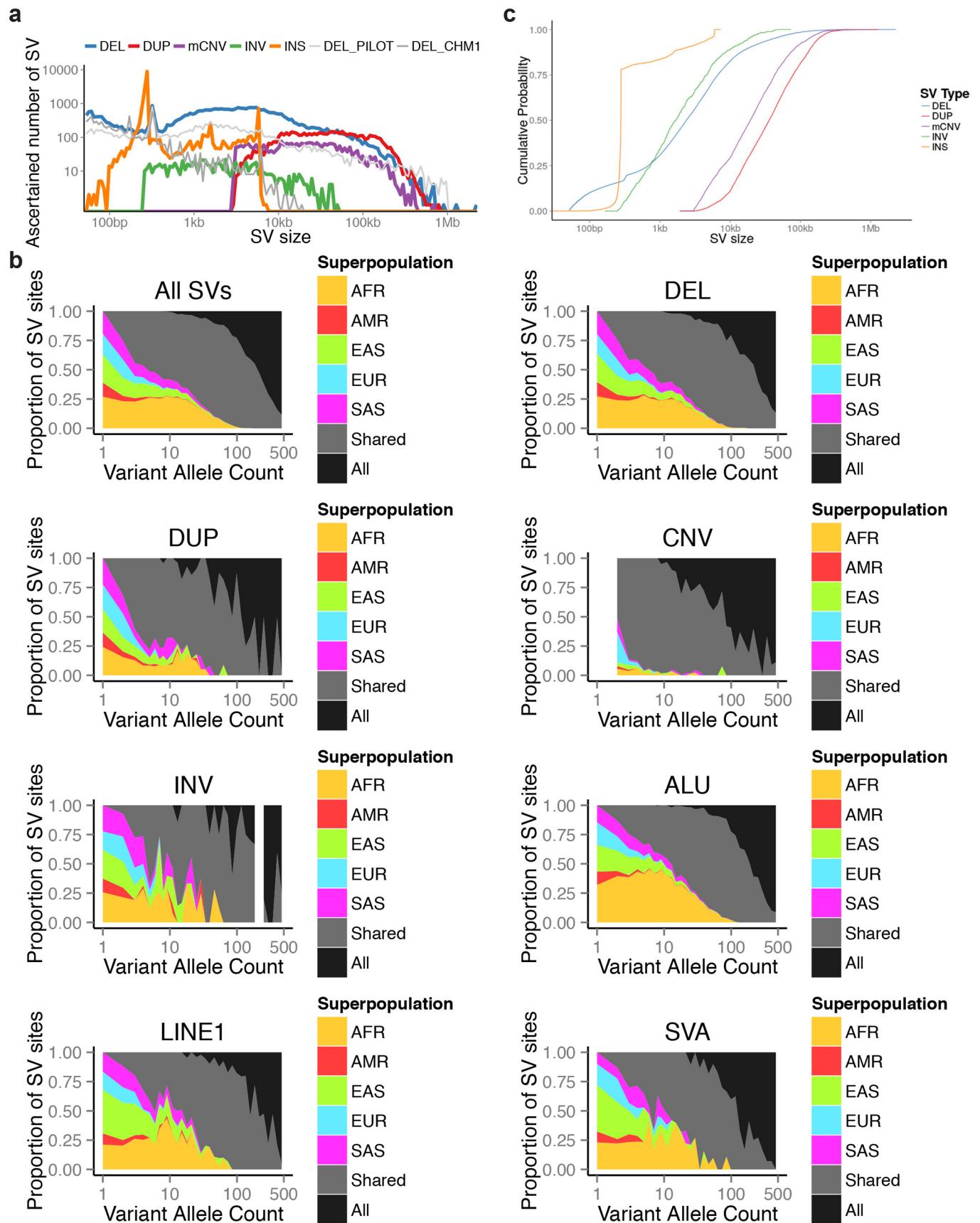
Extended Data Figure 1 | Construction of the SV release and intensity rank sum validation. **a**, Approach used for constructing our SV release set. **b**, Intensity rank sum (IRS) validation results for deletions in different size bins. **c**, IRS validation results for deletions in variant allele frequency (VAF) bins.

d, IRS results for duplications in different size bins. **e**, IRS validation results for duplications in VAF bins. Based on Affymetrix SNP6 array probes, the IRS FDR for all SV length and VAF bins was $\leq 5.4\%$, requiring at least 100 SVs per bin with an IRS assigned P -value.

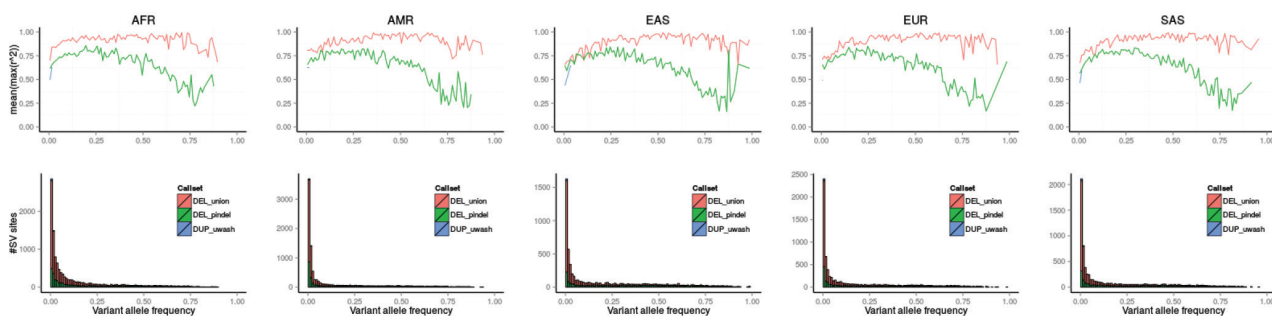
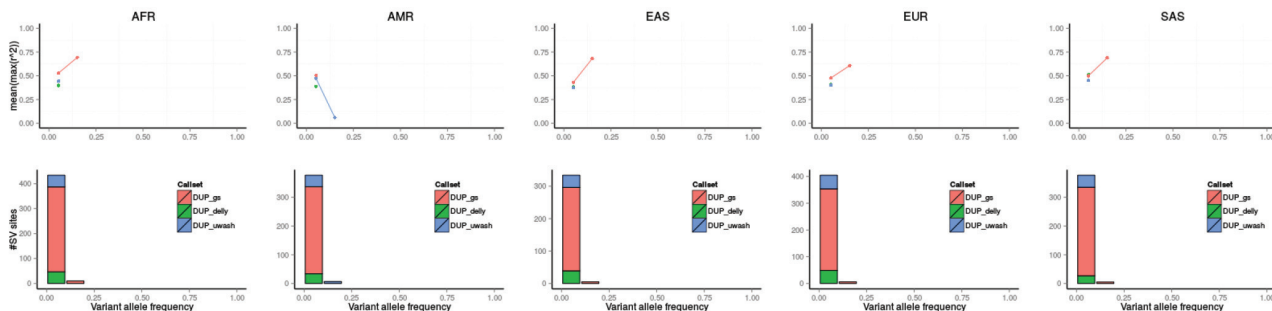
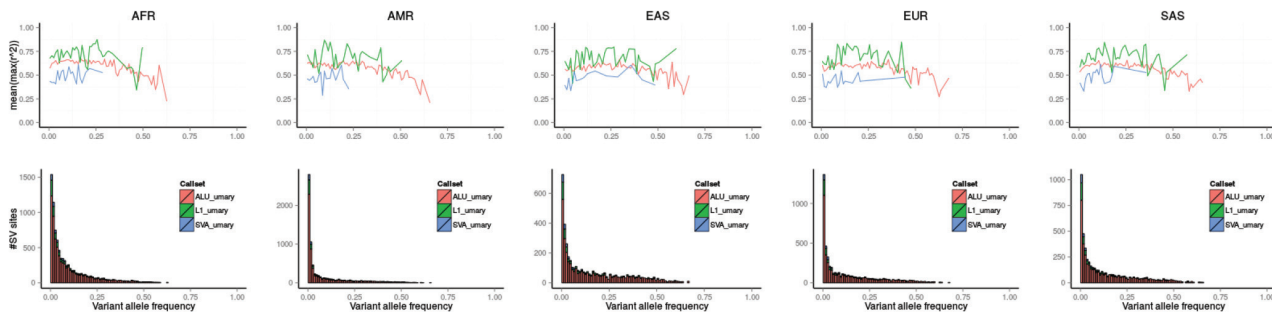
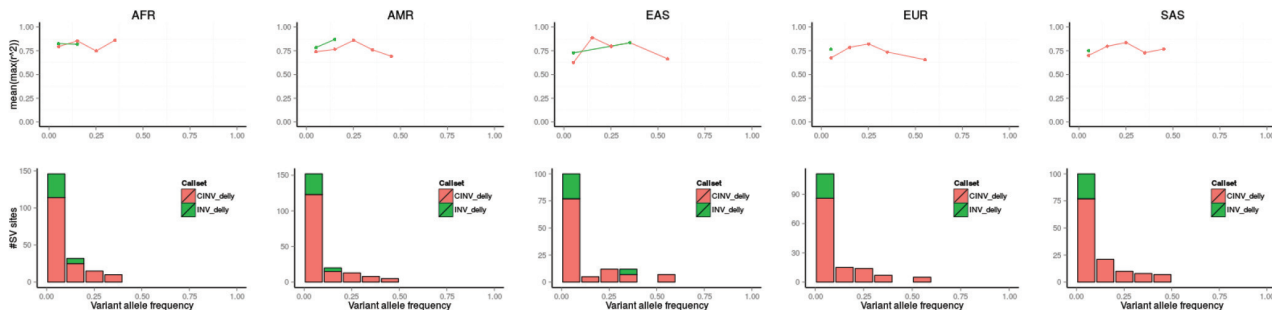


Extended Data Figure 2 | This figure shows the number of SV sites in our phase 3 release relative to allele frequency expressed in terms of allele count. SVs down to an allele count of 1 (corresponding to VAF = 0.0002) are represented in our phase 3 SV set (with the exception of mCNVs, denoted

'CNV' in this figure, which are defined as sites of multi-allelic variation thus requiring allele count ≥ 2 , hence no mCNV sites are ascertained for allele count = 1).

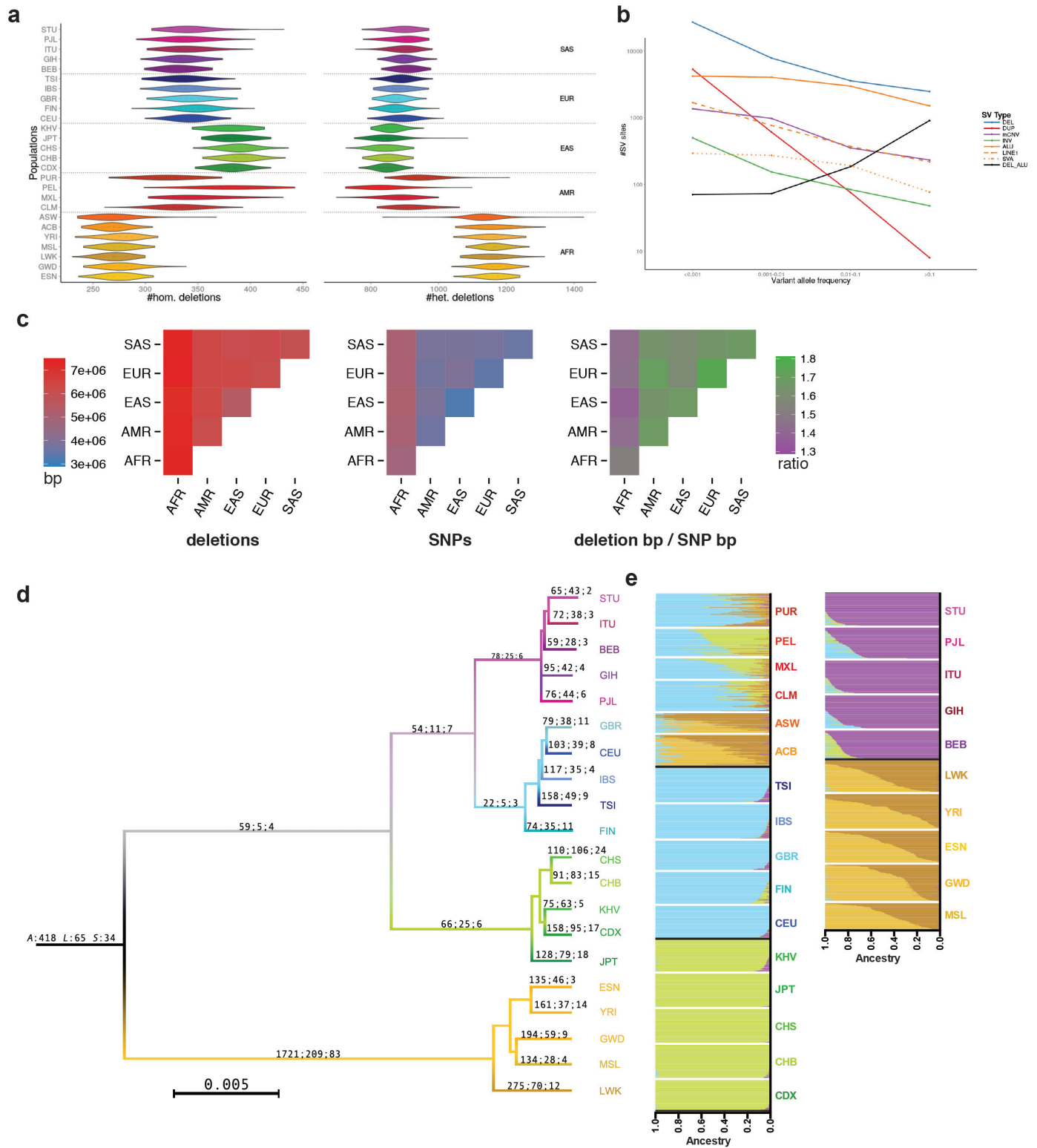


Extended Data Figure 3 | Size and population distribution of different SV classes. **a**, Variants ascertained in the 1000 Genomes Project pilot phase⁶ (light grey) as well as the recent publication of SVs ascertained by PacBio sequencing in the CHM1 genome¹⁴ (grey) are displayed for comparison in this SV size distribution figure (INS, used as abbreviation for MEIs and NUMTs in this display item). **b**, Population distribution of SV allele sharing across continental groups for different SV classes. **c**, Cumulative distributions of the number of events as a function of size by SV class.

a**b****c****d**

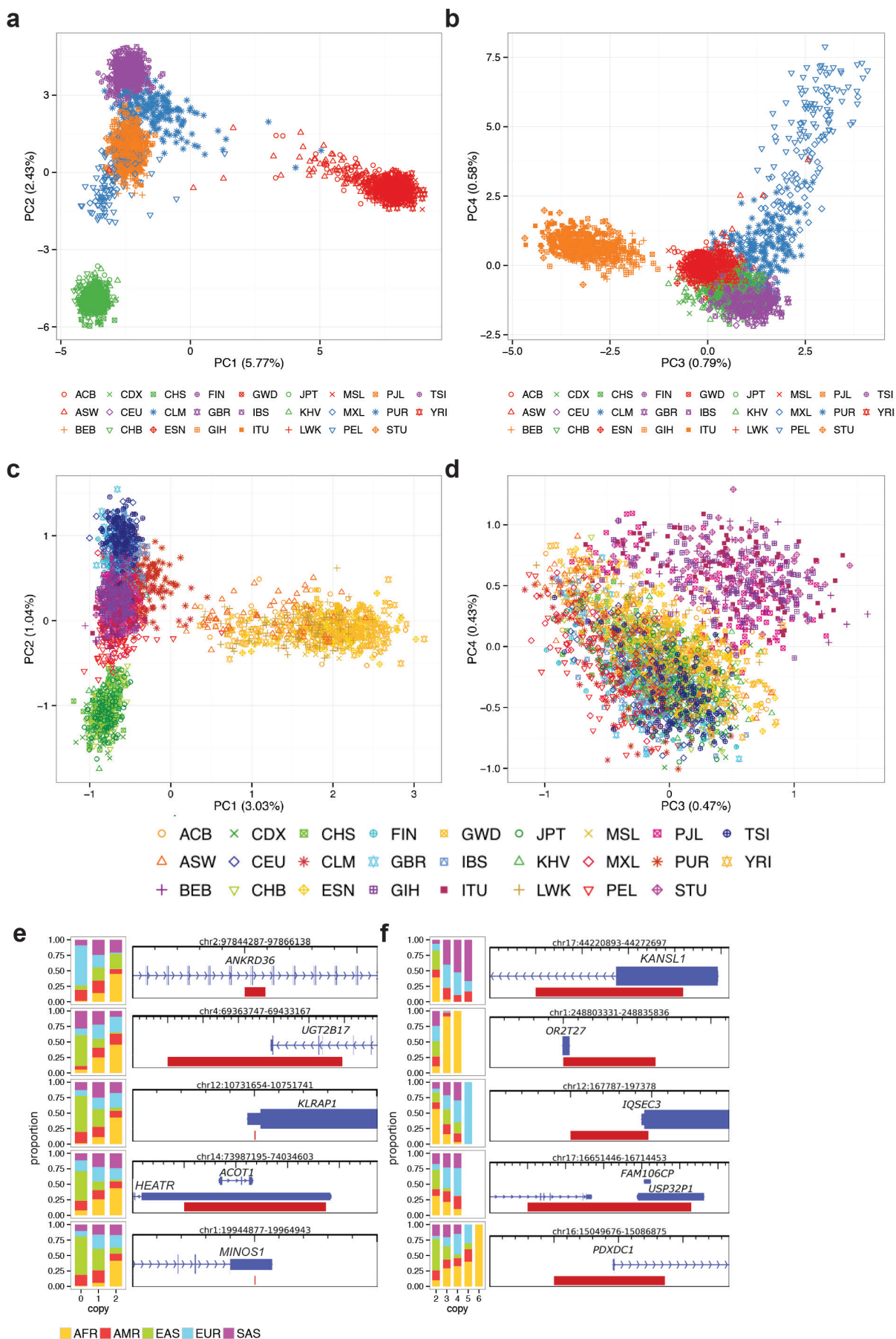
Extended Data Figure 4 | LD properties of various SV classes. **a**, LD properties of deletions, broken down by continental group and shown as a function of VAF. **b**, LD properties of duplications. **c**, LD properties of *Alu*, L1 and SVA mobile element insertions. **d**, LD properties of inversions (with breakdown for two independent inversion sets generated with our inversion

discovery algorithm Delly; that is, CINV = one-sided inversions with support for one breakpoint; INV = two-sided inversions with support for both breakpoints; these two sets are combined into the joint phase3 SV group inversion set).



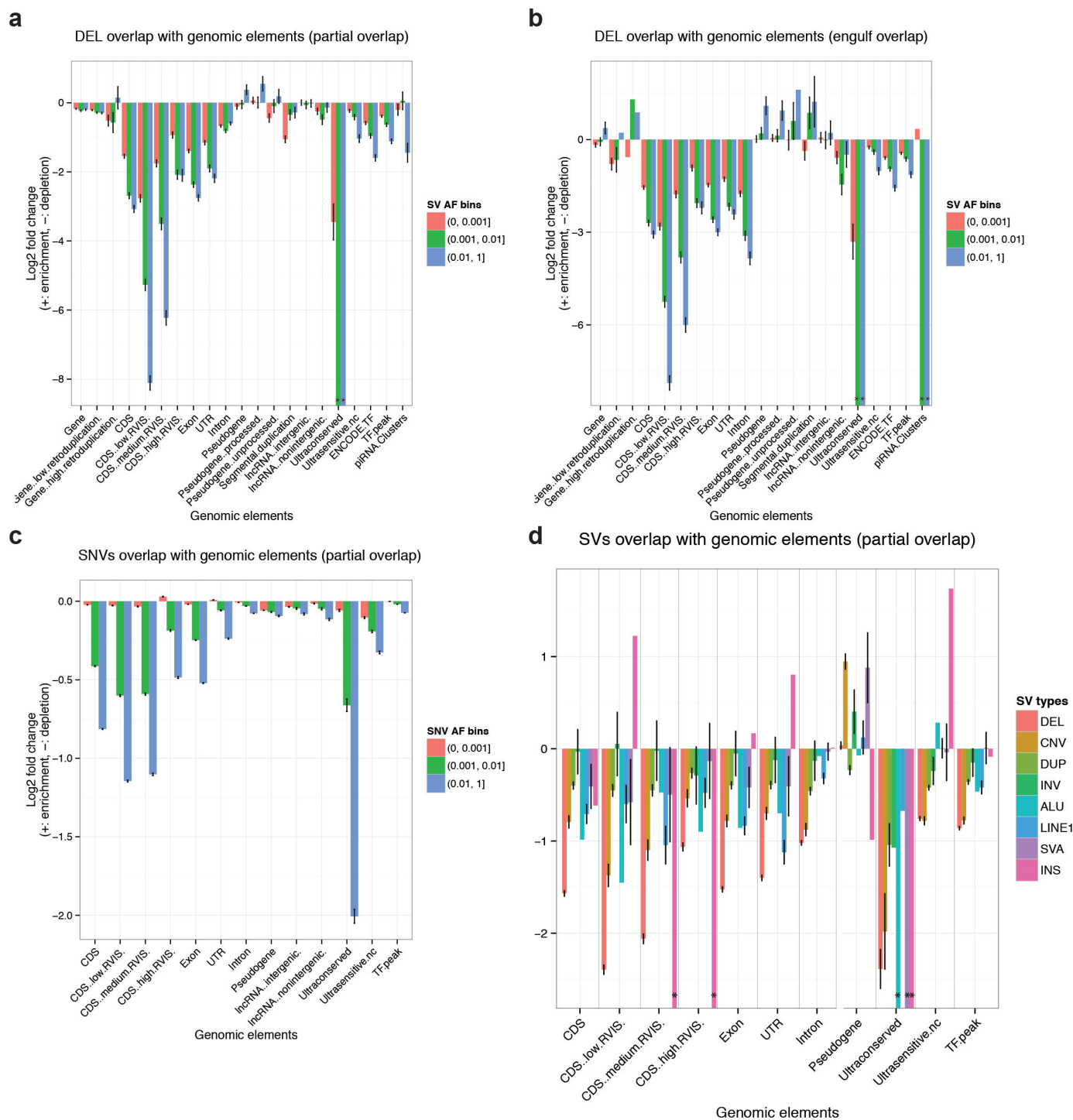
Extended Data Figure 5 | Population genetic properties of SVs. **a**, Deletion heterozygosity and homozygosity among human populations for a subset of high-confidence deletions. Populations from the African continental group (AFR) exhibit the highest levels of heterozygosity and thus diversity among humans, but show the overall lowest level of deletion homozygosity among all continental groups. By comparison, East Asian populations exhibited the lowest levels of deletion heterozygosity and the highest levels of homozygosity. Het., heterozygous; Hom., homozygous. **b**, VAF distribution of major SV classes. Bi-allelic duplications represent a notable outlier, showing a striking depletion of common alleles, which can be explained by the preponderance of genomic sites of duplication to undergo recurrent rearrangement (see main text). As a consequence, most common duplications are classified as multi-allelic variants (that is, mCNVs). **c**, The number of base pairs (bp) differing

among individuals within and between continental groups for deletions (upper panel) and SNPs (middle panel) contrasted with the ratio of deletion bp differences to SNP bp differences ('deletion bp/SNP bp') among groups (lower panel). Non-African groups exhibit a higher 'deletion bp/SNP bp' compared to Africans. **d**, Neighbour-joining tree of populations constructed from MEIs (homoplasy-free markers) to provide a (simplified) view of population ancestry. The tree is labelled with the number of lineage-specific MEIs (*Alu*:L1:SVA). **e**, Classification of ancestry in AFR/AMR and AMR admixed populations using homoplasy-free ancestry informative MEI markers. Colour usage follows the same scheme as in Fig. 1d, except in the case of AFR individuals, which use both the colour in Fig. 1d and another colour that is unrelated to any other figure to indicate additional substructure within this group.



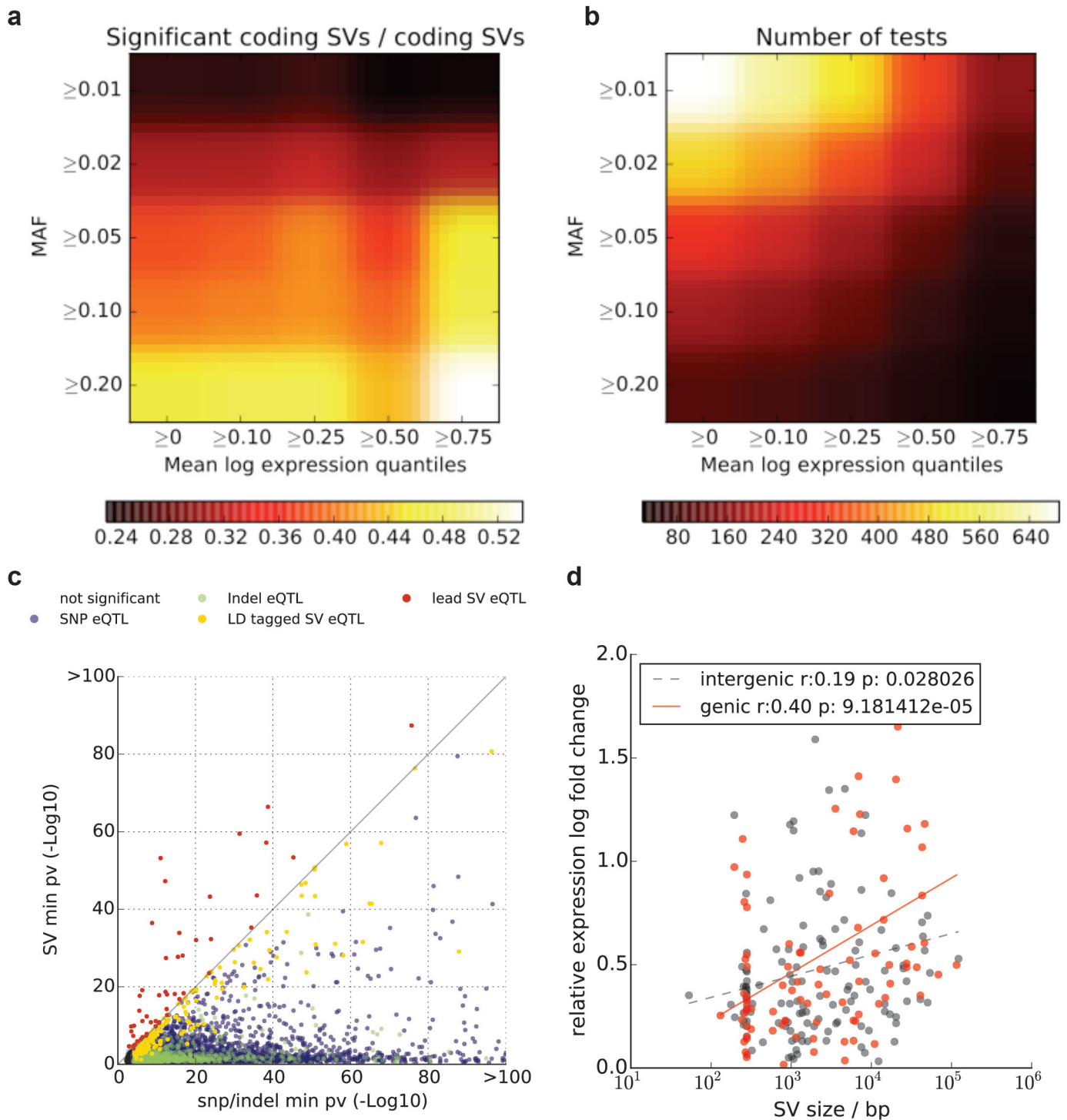
Extended Data Figure 6 | Principal component analysis and population stratification of SVs. **a**, Principal component analysis (PCA) plot of principal components 1 and 2 for deletions. **b**, PCA plot of principal components 3 and 4 for deletions. **c**, PCA plot of principal components 1 and 2 for MEIs. **d**, PCA plot of principal components 3 and 4 for MEIs.

e, The five most highly population-stratified deletions intersecting protein-coding genes based on V_{ST} . **f**, The five most highly population-stratified duplications and multi-allelic copy number variants (mCNVs) intersecting protein-coding genes based on V_{ST} . For abbreviations, see Supplementary Table 1.



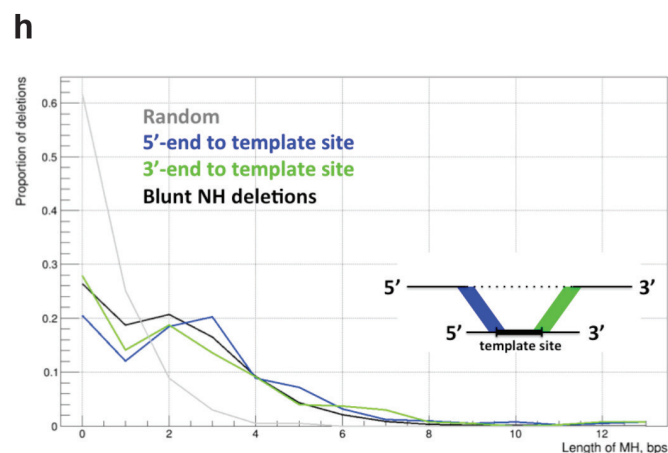
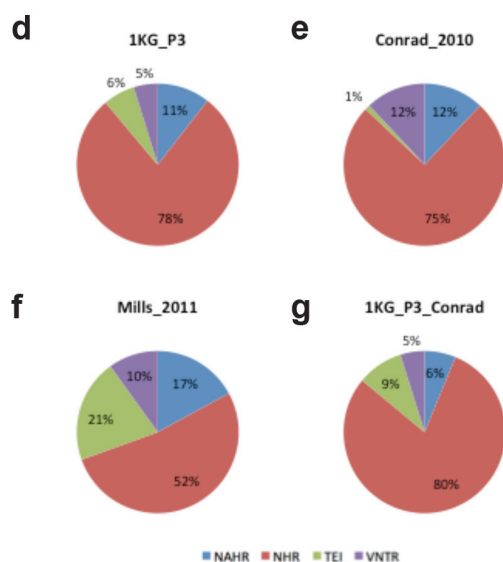
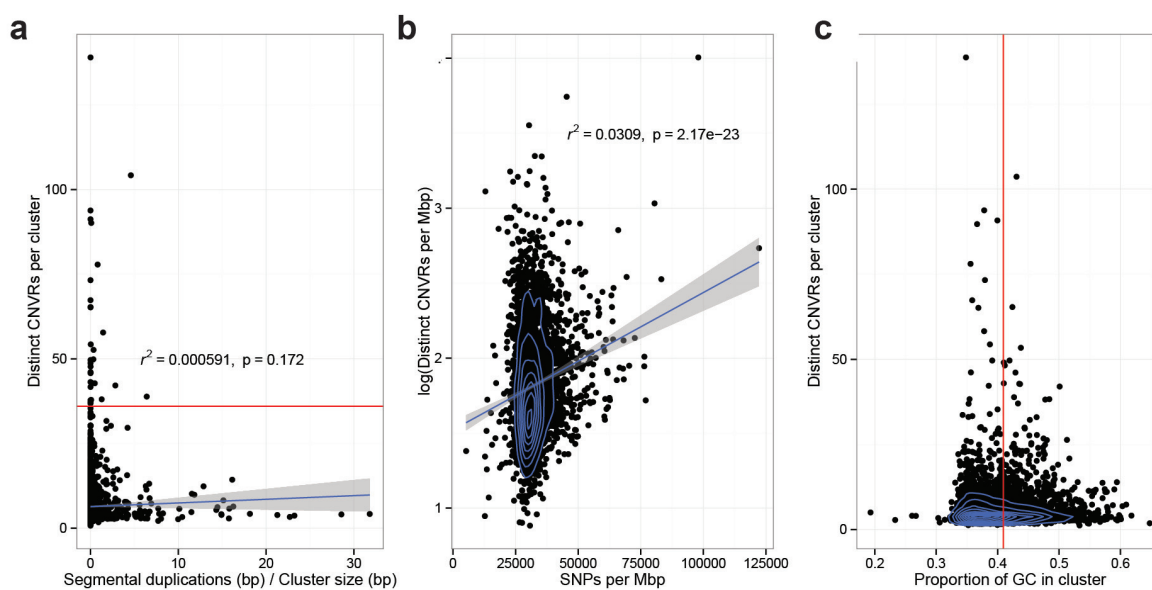
Extended Data Figure 7 | Enrichment of functional elements intersecting SVs. **a**, Shadow figure of Fig. 2a. Overlap enrichment analysis of deletions (with resolved breakpoints) versus genomic elements, using partial overlap statistic, deletions categorized into VAF bins. **b**, Similar to **a**. The only difference is that engulf overlap statistic is used instead of partial overlap statistic. Engulf overlap statistic is the count of genomic elements (for example, CDS) that are fully imbedded in at least one SV interval (for example,

deletions). *no element intersected observed within data set. **c**, Similar to **a** and **b**, with the enrichment/depletion analysis pursued for common SNPs as well as more rare single nucleotide polymorphisms/variants (SNVs). Common SNV alleles show the highest levels of depletion for investigated genomic elements. **d**, Overlap enrichment analysis of various SV types versus genomic elements, using partial overlap statistic.



Extended Data Figure 8 | SV-eQTL analysis. **a**, SV-centric eQTL analysis of coding SVs. Shown is the proportion of coding SVs that are eQTLs as a function of the minimum VAF and the expression quartile. **b**, Total number of coding SVs for corresponding filters. Common SVs ($\text{VAF} > 0.2$) in highly expressed genes ($>75\%$ quantile) are very likely to correspond to SV-eQTLs (54%, see also Supplementary Table 8). **c**, For all genes with significant eQTLs ($\text{FDR} < 10\%$), shown are raw P -values considering only SNPs (x axis) or only SVs (y axis). Genes with (strict lead) SV-eQTLs are shown in red. Genes

with a SNP lead eQTL that is in linkage with an SV ($r^2 > 0.5$) are shown in orange. SNP lead eQTLs without an SV in LD are shown in blue. **d**, Relative eQTL effect sizes for genetic and intergenic SV eQTLs ($n = 239$) either with an SV-eQTL or an LD tagged SV (in log abundance scale). Shown are regression trends for both genic and intergenic SV eQTLs. For genetic eQTLs, a clear relationship between SV effect size is found. For example, genic SVs >10 kb have threefold larger effect sizes compared to genic SVs < 1 kb; $P = 0.004$; t -test.

[illegible]

Extended Data Figure 9 | SV clustering and breakpoint analysis. a–c,

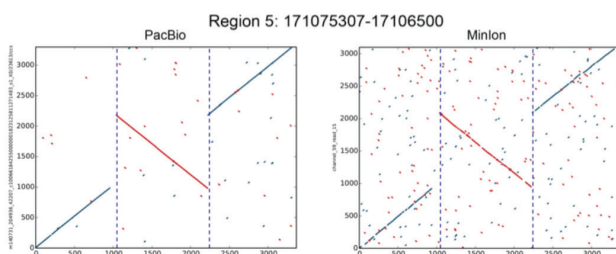
Extensive clustering of recurrent SVs into CNVRs appears unrelated to the extent of segmental duplications (a) and is only partially correlating with SNP diversity (b) and GC content (c). Breakdown of SV mechanism classifications based on criteria from two earlier studies (refs 6, 40). Shown are results for deletions with nucleotide resolved breakpoints. BreakSeq was used for mechanism inference. d, 1KG_P3: breakdown for our 1000 Genomes Project phase 3 SV callset using classification criteria from ref. 6. e, Conrad_2010: summary of mechanism classification results published in ref. 40. f, Mills_2011: summary of mechanism classification results published in ref. 6.

g, 1KG_P3_Conrad: Breakdown for our 1000 Genomes Project phase 3 SV callset using classification criteria from ref. 40. Mechanism classification was pursued using four different categories. Blue, non-allelic homologous recombination (NAHR); green, mobile elements inserted into the reference genomes (appearing deleted in this analysis); red, non-homology-based

rearrangement mechanisms (NHR), such as NHEJ, microhomology-mediated end-joining and microhomology-mediated break-induced replication (involving blunt-ended deletion breakpoints or breakpoints with microhomology); purple, expansion or shrinkage of variable numbers of tandem repeats (VNTRs). TEI, transposable element insertion (equivalent with MEI). h, Distribution of lengths of micro-homology (MH) for complex SVs, measured between deletion and corresponding template sites boundaries. Simple deletions, which based on BreakSeq were inferred to be formed by a non-homology-based SV formation mechanism, such as NHEJ and microhomology-mediated break-induced replication (Supplementary Table 3), are shown as an additional control (here denoted 'blunt NH deletions'). i, Origins of inserted sequences in complex deletions inferred by split read analysis. This figure depicts examples for each class shown in Supplementary Table 13.

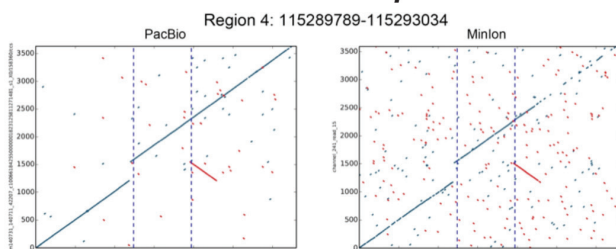
Class 1 - simple inv

a



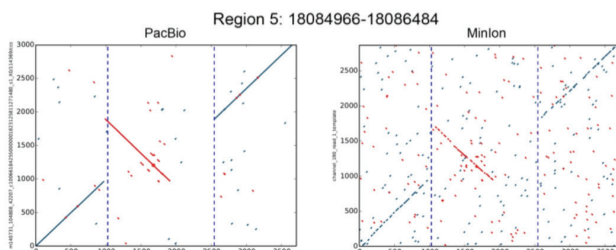
Class 2 - inv-dup

b



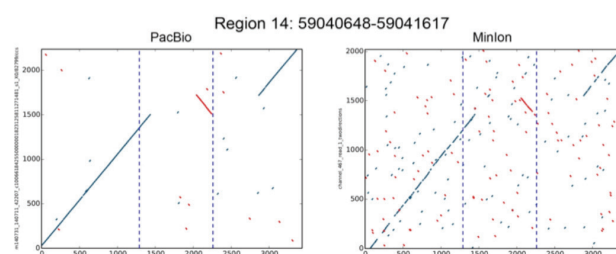
Class 3 - inv-del

C



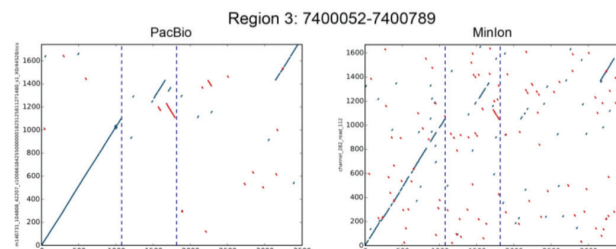
Class 4 - inv-2dels

d

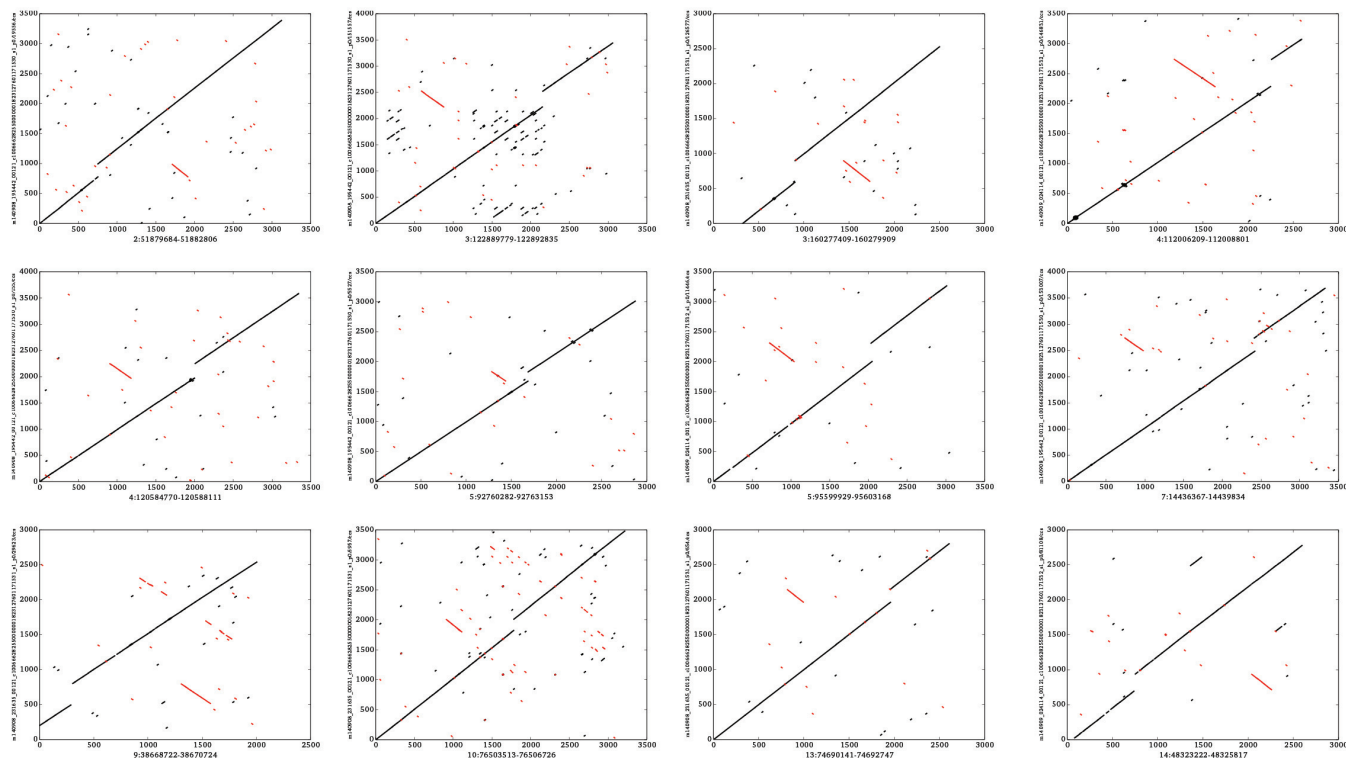


Class 5 - complex

e



f



Extended Data Figure 10 | Examples of inversions identified in the SV

release. **a–e**, Five classifications of inversions verified using PacBio and Minion reads are represented: Simple Inversion (**a**), inv-dup (**b**), inv-del (**c**), MultiDel with Inv (here abbreviated as inv-2dels) (**d**) and complex (**e**). **f**, Several further examples of inverted duplications (inv-dup), the most common form of inversion-associated SV identified in the phase 3 release set. The figure is

depicting DNA sequence alignment dotplots (same arrangement as in Fig. 3), with the *y* axis referring to PacBio DNA single molecule sequencing reads and the *x* axis referring to the reference genome assembly (hg19). Inverted sequences are highlighted in red. Sequence analysis suggests that these inverted duplications are not typically associated with retrotransposition.

The UK10K project identifies rare variants in health and disease

The UK10K Consortium*

The contribution of rare and low-frequency variants to human traits is largely unexplored. Here we describe insights from sequencing whole genomes (low read depth, 7×) or exomes (high read depth, 80×) of nearly 10,000 individuals from population-based and disease collections. In extensively phenotyped cohorts we characterize over 24 million novel sequence variants, generate a highly accurate imputation reference panel and identify novel alleles associated with levels of triglycerides (*APOB*), adiponectin (*ADIPOQ*) and low-density lipoprotein cholesterol (*LDLR* and *RGAG1*) from single-marker and rare variant aggregation tests. We describe population structure and functional annotation of rare and low-frequency variants, use the data to estimate the benefits of sequencing for association studies, and summarize lessons from disease-specific collections. Finally, we make available an extensive resource, including individual-level genetic and phenotypic data and web-based tools to facilitate the exploration of association results.

Assessment of the contribution of rare genetic variation to many human traits is still largely incomplete. In common and complex diseases, a lack of empirical data has to date hampered the systematic assessment of the contribution of rare and low-frequency genetic variants (defined throughout this paper as minor allele frequency (MAF) <1% and 1–5%, respectively). Rare variants are incompletely represented in genome-wide association (GWA) studies¹ and custom genotyping arrays^{2,3}, and impute poorly with current reference panels. Rare and low-frequency variants also tend to be population- or sample-specific, requiring direct ascertainment through resequencing^{4,5}. Recent exome-wide resequencing studies have begun to explore the contribution of rare coding variants to complex traits⁶, but comparatively little is known of the non-coding part of the genome where most complex trait-associated loci lie⁷. At the other end of the human disease spectrum, the widespread application of exome-wide sequencing is accelerating the rate at which genes and variants causal for rare diseases are being identified. Despite this, many Mendelian diseases still lack a genetic diagnosis and the penetrance of apparently disease-causing loci remains inadequately assessed.

The UK10K project was designed to characterize rare and low-frequency variation in the UK population, and study its contribution to a broad spectrum of biomedically relevant quantitative traits and diseases with different predicted genetic architectures. Here we describe the data and initial findings generated by the different arms of the UK10K project. In addition to this paper, UK10K companion papers describe the utility of this resource for imputation⁸, association discovery for bone mineral density⁹, thyroid function¹⁰ and circulating lipid levels¹¹ and provide access to the study results through novel web tools¹².

Study designs in the UK10K project

The UK10K project includes two main project arms (Table 1). The UK10K-cohorts arm aimed to assess the contribution of genome-wide genetic variation to a range of quantitative traits in 3,781 healthy individuals from two intensively studied British cohorts of European ancestry, namely the Avon Longitudinal Study of Parents and Children (ALSPAC)¹³ and TwinsUK¹⁴. A low read depth (average 7×) whole-genome sequencing (WGS) strategy was employed in

order to maximize total variation detected for a given total sequence quantity¹⁵ while allowing interrogation of noncoding variation. Sixty-four different phenotypes were analysed, including traits of primary clinical relevance in 11 major phenotypic groups (obesity, diabetes, cardiovascular and blood biochemistry, blood pressure, dynamic measurements of ageing, birth, heart, lung, liver and renal function; Supplementary Table 1). Of these, 31 phenotypes were available in both studies (referred to as 'core' and reported in association analyses), 18 were unique to TwinsUK and 15 were unique to ALSPAC.

The UK10K-exomes arm aimed to identify causal mutations through high read depth (mean ~80× across studies) whole-exome sequencing of approximately 6,000 individuals from three different collections: rare disease, severe obesity and neurodevelopmental disorders. The disorders studied in the UK10K-exomes arm have been shown to have a substantial genetic component at least partially driven by very rare, highly penetrant coding mutations. The rare disease collection includes 125 patients and family members in each of eight rare disease areas (Table 1). Disease types were selected with different degrees of locus heterogeneity, prior evidence for monogenic causation and likely modes of inheritance (for example, dominant or recessive). The obesity collection comprises of samples with severe obesity phenotypes, including approximately 1,000 subjects from the Severe Childhood Onset Obesity Project (SCOOP)¹⁶, plus severely obese adults from several population cohorts. The neurodevelopmental collection comprises of ~3,000 individuals selected to study two related neuropsychiatric disorders (autism spectrum disorder and schizophrenia).

Discovery of 24 million novel genetic variants

In total, 3,781 individuals were successfully whole-genome sequenced in the UK10K-cohorts arm. After conservative quality control filtering (Extended Data Figs 1 and 2 and Supplementary Table 2), the final call set contained over 42M single nucleotide variants (SNVs, 34.2M rare and 2.2M low-frequency), ~3.5M insertion/deletion polymorphisms (INDELs; 2,291,553 rare and 415,735 low-frequency) and 18,739 large deletions (median size 3.7 kilobase). Each individual on average contained 3,222,597 SNVs (5,073 private), 705,684 INDELs (295 private) and 215 large deletions (less than 1 private). Of 18,903 analysed

*A list of authors and affiliations appears at the end of the manuscript.

protein-coding genes, 576 genes contained at least one homozygous or compound heterozygous variant predicted to result in the loss of function of a protein (LoF, Supplementary Information, 14,516 variants in total). As previously shown^{5,17}, variants predicted to have the greatest phenotypic impact (LoF and missense variants, and variants mapping to conserved regions), were depleted at the common end of the derived allele spectrum (Extended Data Fig. 3). There were 495 homozygous LoF variants, a subset of which associated with phenotypic outliers (Supplementary Table 3).

We assessed sequence data quality by comparison with an exome sequencing data set (WES, $\sim 50\times$ coverage)¹⁸ and in 22 pairs of monozygotic twins (Extended Data Fig. 1). The non-reference discordance (NRD, or the fraction of discordant genotypes for non-reference homozygous or heterozygous alleles) was 0.6% for common variants and 3.2% (range 0.1–3.3%; Extended Data Fig. 1) for low-frequency and rare variants. False discovery rates (FDR) were comparable between newly discovered sites and sites previously reported in the 1000 Genomes Project phase 1 (1000GP) data set⁵.

When compared to two large-scale European sequencing repositories, 1000GP and the Genome of the Netherlands (GoNL, $12\times$ read depth¹⁹), UK10K-cohorts discovered over 24M novel SNVs. Overall, 96.5% of variants with MAF $> 1\%$ were shared, reflecting a common reservoir within Europe (Fig. 1 and Extended Data Fig. 2). Conversely, 94.7% of singleton (allele count (AC) = 1) and 55.0% of rare (AC > 1 and MAF $< 1\%$) SNVs were study-specific. In a similar comparison, 64.4% (AC = 1) and 15.8% of variants (AC > 1 and MAF $< 1\%$) found in GoNL were found to be study-specific compared to 1.2% of variants above 1% MAF.

This deeper characterization of European genetic and haplotype diversity will benefit future studies by creating a novel genotype imputation panel with substantially increased coverage and accuracy compared to the 1000GP reference panel⁸ (see ref. 9 and the next section for its application). It further informs a detailed empirical assessment of the geographical structure of rare variation in the UK where we detected geographical structure for very rare alleles (AC = 2–7) in Northern and Western UK regions, although this did not show evidence of substantial correlation with variation in phenotype (Box 1).

Findings from single-marker association tests

A main aim of the UK10K-cohorts project was to assess associations of low-frequency and rare variants under different analytical strategies (Fig. 2). We used a unified analysis strategy for the parallel evaluation of all quantitative traits (Supplementary Information, Supplementary Table 4). Here we describe results for the 31 core traits shared in ALSPAC and TwinsUK, with other results reported elsewhere¹².

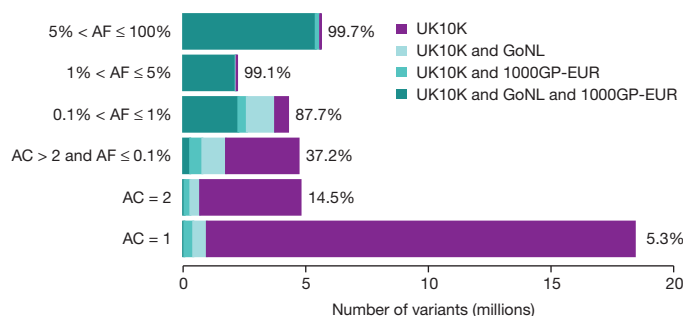


Figure 1 | The UK10K-cohorts resource for variation discovery. Number of SNVs identified in the UK10K-cohorts data set in all autosomal regions in different allele frequency (AF) bins, and percentages that were shared with samples of European ancestry from the 1000 Genomes Project (phase I, EUR $n = 379$) and/or the Genomes of the Netherlands (GoNL, $n = 499$) study, or unique to the UK10K-cohorts data set. AF bins were calculated using the UK10K data set, for allele count (AC) = 1, AC = 2, and non-overlapping AF bins for higher AC. All numerical values are in Extended Data Fig. 2.

We first carried out single-marker association tests, as in standard genome-wide association studies of common variants²⁰. Assuming an additive genetic model, we used standard approaches to model relationships between standardised traits, residualized for relevant covariates, and allele dosages of 13,074,236 SNVs, 1,122,542 biallelic INDELs (MAF $\geq 0.1\%$) and 18,739 large deletions in whole-genome sequenced samples ('WGS sample'). We further assessed associations in an independent study sample of genome-wide genotyped individuals ('GWA' sample) including up to 6,557 ALSPAC and 2,575 TwinsUK participants who were not part of UK10K (actual numbers per trait are given in Supplementary Table 1). In the GWA sample, genotypes were imputed from genome-wide single nucleotide polymorphism (SNP) data using the UK10K haplotype reference panel, described in a companion manuscript⁸. The combined WGS+GWA sample had 80% power to detect associations of SNVs of low-frequency and rare down to \sim MAF 0.5%, for a per-alleles trait change (the regression beta coefficient or Beta) of ~ 1.2 standard deviations or greater (Fig. 3). To combine WGS and GWA data we carried out a fixed effect meta-analysis using the inverse variance method, which showed no evidence of inflation of summary statistics at the traits investigated (GC lambda ≈ 1). We used a conservative stepwise procedure for reporting loci from single-variant analysis (Supplementary Table 5), and we discuss elsewhere replication and technical validation of associations of rare variants not supported in the combined WGS+GWA sample (Supplementary Information, Supplementary Table 6).

Table 1 | Summary of sample collections and sequencing metrics for the four main studies of the UK10K project

| Study name and design | <i>n</i> | Sequencing strategy, mean read depth and Ts/Tv ratio | SNVs/INDELs | SNVs/INDELs by allele frequency |
|--|---------------|--|----------------------|--|
| Cohorts. Unselected samples from two population-based cohorts | 3,781 | WGS, $7\times$ Ts/Tv = 2.15 | 42,001,210/3,490,825 | <1%: 34,247,969/2,296,962 1–5%: 2,298,220/412,168 >5%: 5,869,317/1,496,955 |
| Rare. Eight rare diseases with expected different allelic architectures (ciliopathy, coloboma, congenital heart disease, familial hypercholesterolaemia, intellectual disability, neuromuscular, severe insulin resistance and thyroid disease) | 961 (397) | WES, $77\times$ Ts/Tv = 3.02 | 252,809/1,621 | <1%: 171,564/1,384 $\geq 1\%$: 81,245/237 |
| Obesity. Severely obese children (BMI > 3 s.d. from population mean) and adults with extreme obesity | 1,468 (1,359) | WES, $82\times$ Ts/Tv = 3.02 | 484,931/3,370 | <1%: 403,684/3,133 $\geq 1\%$: 81,247/237 |
| Neurodevelopmental. Autism and schizophrenia (individual probands, families with one affected and other healthy individuals sampled, families with data from multiple affected individuals and individuals with comorbid intellectual disability and psychosis) | 2,753 (1,707) | WES, $77\times$ Ts/Tv = 3.02 | 538,526/3,826 | <1%: 457,278/3,589 $\geq 1\%$: 81,248/237 |

For the cohorts arm, numbers are for the set of 3,781 samples passing quality control, while a subset of 3,621 was used for association testing. For the exome arm, numbers of sites are based on the joint call set, and are calculated for a subset of all individuals that represent the patient subset (in brackets). The total number of individuals sequenced in each study is also given (see Supplementary Methods). The transition to transversion ratio (Ts/Tv) was calculated for the final set of SNVs excluding multiallelic sites. WGS, whole-genome sequencing; WES, whole-exome sequencing.

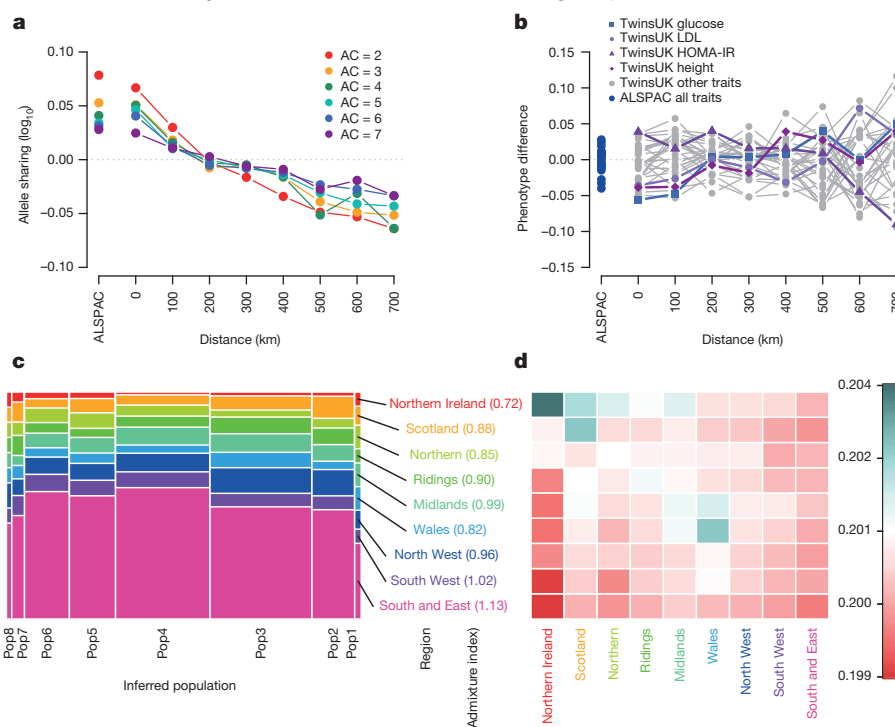
BOX 1

Genetic structure of rare variation within the UK

We used the ALSPAC cohort (from the Bristol region) and a subset of TwinsUK individuals (UK-wide origin) to investigate the spatial structure of rare genetic variants (Supplementary Table 16). We first sought to define the extent to which variants of different MAF were geographically structured. We estimated the excess of allele sharing between pairs of individuals as a function of their physical distance, as compared to expectations under a neutral model (Supplementary Information)⁴⁶. Rare genetic variants showed excess allele sharing at distances smaller than about 200 km, and reduced sharing for more than about 300 km. There was a steeper geographical cline for doubletons ($AC = 2$), which decreased with increasing allele counts (3 up to 7, equivalent to a MAF of ~ 0.1 – 0.3% ; **a**). No corresponding geographical structure was observed for phenotypic variation (**b**).

We next assessed the extent to which the non-random distribution of rare SNVs could be accounted for by regional differences at the level of 13 main regions within the UK⁴⁷. Overall, patterns of allele sharing were indicative of a larger degree of genetic homogeneity in Southern and Eastern England compared to individuals of Welsh, Northern, Scottish or Northern Irish origin. Doubletons were the most structured both within and between regions (Wilcoxon rank sum P value < 0.05 , Extended Data Fig. 8).

Finally, we used “chromosome painting”⁴⁸ to gain insights into possible demographic events underlying the observed genetic structure. We first estimated the average length of DNA tracts shared between individuals, and used the number of such tracts to identify fine population structure in our data set. The tract length distribution showed weak geographic structure reflecting the rare variant analysis. A fine structure analysis suggested that the identified populations were not strongly geographically defined, indicative of a large degree of movement between regions compared to the samples in the Peoples of the British Isles study⁴⁵, which were chosen to have all four grandparents born in the same location (Extended Data Fig. 9).



Box 1 | Population structure in UK10K-cohorts. All ALSPAC (from Bristol), and 1,139 TwinsUK (UK-wide) participants with a complete set of genotype, phenotype and place of birth data. **a**, Excess of allele sharing as a function of geographical distance, expressed as the proportion of shared alleles between sample pairs for AC from 2 to 7 against their geographical distance. **b**, Phenotypic sharing, estimated for the 31 core phenotypes as the absolute difference between pairs of individuals, averaged within distance bins, rescaled and plotted against their geographical distance. The four traits with the most extreme structure are highlighted. HOMA-IR, homeostatic model assessment for insulin. **c**, Geographical decomposition of each population. Populations are shown proportional to size; historically ‘Celtic’ and ‘Briton’ regions are closer to the edges, whereas ‘Anglo-Saxon’ England is more homogeneous and at the centre (see ref. 45). Ridings refers to East and West Ridings, Yorkshire. **d**, Average length of DNA tracts shared between individuals when clustered by sampling location. The ‘admixture’ index is given in brackets, with one-third corresponding to regions containing completely unadmixed populations and infinity to completely admixed populations. See also Extended Data Fig. 9.

Overall, across the 31 traits 27 independent loci reached our experiment-wide significance threshold²¹ P value $\leq 4.62 \times 10^{-10}$ in the combined WGS+GWA sample (Fig. 3 and Supplementary Table 5). Two associations have been newly discovered by this project, and were conditionally independent of other variants previously reported at the same loci. The first was a low-frequency intronic variant in *ADIPOQ* associated with decreased adiponectin levels (rs74577862-A, effect allele frequency (EAF) = 2.6%, P value = 3.04×10^{-64}). The second was a rare splice variant (rs138326449) in *APOC3* described in advance of this manuscript^{11,22,23}. The remaining 25 loci reaching experiment-wide significance in the combined

WGS+GWA sample included common, low-frequency and rare variants tagging known associations with adiponectin levels (*CDH13* and *ADIPOQ*), lipid traits (*APOB*, *APOC3-APOA1*, *APOE*, *CETP*, *LIPC*, *LPL*, *PCSK9*, *SORT1-PSRC1-CELSR2*), C-reactive protein (*LEPR*), haemoglobin levels (*HFE*) and fasting glycaemic traits (*G6PC2-ABCB11*, Supplementary Table 5). In contrast to previous projections²⁴, from this analysis of a wide range of biomedical traits there was no evidence of low-frequency alleles with large effects upon traits (Fig. 3)²⁵, with classical lipid alleles identifying extremes of single-variant genetic contributions for these traits. This suggests that few, if any, low-frequency variants with stronger effects than those we see

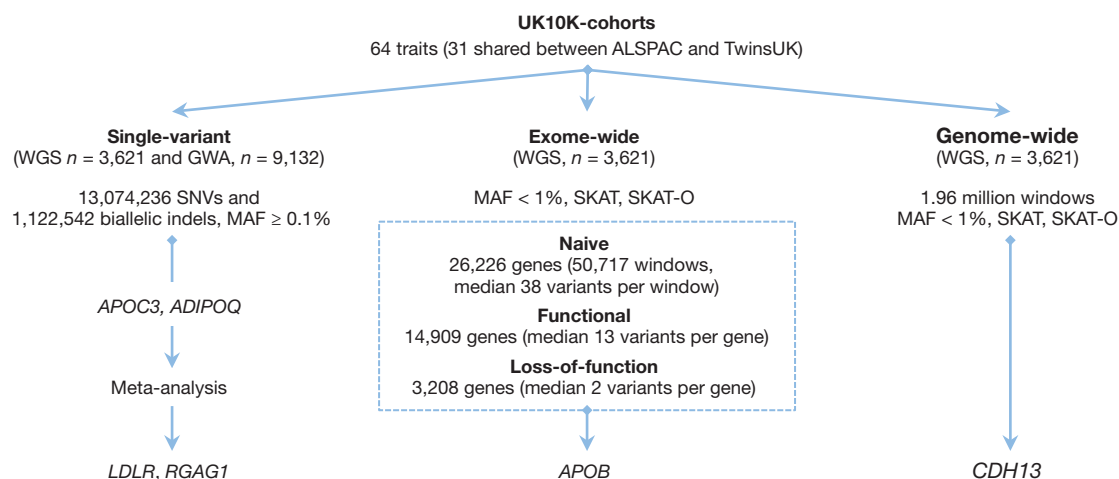


Figure 2 | Study design for associations tested in the UK10K-cohorts study. Summary of phenotype–genotype association testing strategies employed in the UK10K-cohorts study.

are likely to be detected in the general European population for the wide range of traits that we considered.

Increasing sample size may identify additional moderate effect variants, or variants with rarer frequency. We therefore sought to assess the extent to which the more accurate imputation offered by the UK10K reference panel, applied to larger study samples, could discover additional associations. A restricted maximum likelihood (REML)²⁶ analysis suggested that using the UK10K data could increase the estimated variance explained, compared to the sparser HapMap2, HapMap3 and 1000GP data sets (Extended Data Table 1). We tested four lipid traits (high-density and low-density lipoprotein cholesterol, total cholesterol and triglycerides) in up to 22,082 additional samples from 14 cohorts imputed to the combined UK10K+1000GP phase I panel (Supplementary Table 7).

This effort identified two novel associations with low-density lipoprotein cholesterol (Fig. 3, Supplementary Table 8), which we further replicated in an independent imputation data set of 15,586 samples from 8 cohorts and through genotyping in 95,067 samples from the

Copenhagen General Population Study (CGPS²⁷). The first was a rare intronic variant in *LDLR* (rs72658867-A, c.2140 + 5G > A; EAF = 0.01, combined sample *P* value = 1.27×10^{-46}); per allele effect Beta (s.e.m.) = $-0.23 \text{ mmol l}^{-1}$ (0.02), *P* value = 7.63×10^{-30} (CGPS, *n* = 95,079). The second was a common, X-linked variant near *RGAG1* (rs5985471-T, EAF = 0.403, *P* value = 1.53×10^{-12}); per allele effect Beta (s.e.m.) = $-0.02 \text{ mmol l}^{-1}$ (0.004), *P* value = 1.8×10^{-5} (CGPS, *n* = 93,639). The *LDLR* variant was previously classified to be of uncertain impact in ClinVar, and reported to have no effect on plasma cholesterol levels in a small sample of familial hypercholesterolaemia patients²⁸. The *LDLR*-A allele is almost perfectly imputed in our sample (info = 0.96), but absent in previous imputation panels²⁹; the *RGAG1*-T allele is common but was missed in previous studies, which focused predominantly on autosomal variation²⁹. Within CGPS, these variants were weakly associated with ischaemic heart disease (odds ratio (OR) = 0.77(0.66, 0.92), *P* = 0.003 for rs72658867; 0.96(0.94, 0.99), *P* = 0.005 for rs5985471) and rs72658867 with myocardial infarction (OR = 0.65(0.49, 0.87), *P* = 0.003; Supplementary Table 8). These results demonstrate the value of our expanded haplotype reference panel for discovery of trait associations driven by low-frequency and rare variants, as also shown in refs 9, 10.

Findings from rare variant association tests

Single-marker association tests are typically underpowered for rare variants³⁰. Many questions remain regarding the optimal choice of test, owing to the unknown allelic architecture of rare variant contribution to traits, in particular outside protein-coding regions. We first evaluated associations by considering genes (GENCODE v15) as functional units of analysis using three separate variant selection strategies. Naive tests considered all variants in exons, untranslated regions (UTRs) and essential splice sites, weighted equally. Functional tests considered missense and LoF variants, the latter defined as being predicted to cause essential splice site changes, stop codon gains or frameshifts. For each scenario we applied two separate statistical models with different properties, sequence kernel association tests (SKAT) and burden tests implemented in SKAT and SKAT-O^{31,32}, to rare variants (MAF < 1%).

Overall, there was an excess of test statistics with *P* values $\leq 10^{-4}$ for functional and loss-of-function tests (Extended Data Figs 4 and 5), with a total of 9, 70 and 196 genes associated with the 31 core traits with the LoF, functional and naive tests, respectively (Supplementary Table 9). A signal driven by loss-of-function variants in the *APOB* gene (encoding apolipoprotein B) achieved our threshold for experiment-wide significance (*P* value $\leq 1.97 \times 10^{-7}$), in a burden-type test (min *P* value for TG = 7.02×10^{-9}). Overall, 3 singleton LoF variants

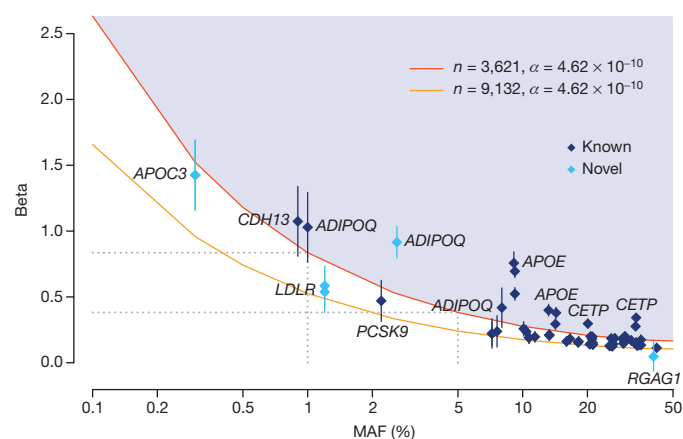


Figure 3 | Summary of association results across the UK10K-cohorts study. Allelic spectrum for single-marker association results for independent variants identified in the single-variant analysis (Supplementary Table 5). A variant's effect (absolute value of Beta, expressed in standard deviation units) is given as a function of minor allele frequency (MAF, x axis). Error bars are proportional to the standard error of the beta, variants identifying known loci are dark blue and variants identifying novel signals replicated in independent studies are coloured in light blue. The red and orange lines indicate 80% power at experiment-wide significance level (*t*-test; *P* value $\leq 4.62 \times 10^{-10}$) for the maximum theoretical sample size for the WGS sample and WGS+GWA, respectively.

were responsible for this signal, of which two were not previously reported (rs141422999 and Chr2:21260958). Examples of novel rare variants in complex trait-associated loci (for example, *G6PC2* associated with fasting glucose) were also seen for genes reaching suggestive levels of association (P value $\leq 10^{-4}$). Lastly, we tested the value of a genome-wide naive approach to explore associations outside protein-coding genes by combining variants across ~ 1.8 million genome-wide tiled windows of 3 kb in size (median 37 SNVs per window, $\text{MAF} < 1\%$, assigning an equal weight to all variants in the window). Overall association statistics appeared underpowered to detect true signals, apart from an association signal for adiponectin driven by a known rare intronic variant at the *CDH13* locus (rs12051272, $\text{EAF} = 0.09\%$, P value $= 6.52 \times 10^{-12}$; Supplementary Table 10)^{33,34}. As previously shown for single-variant tests, in this study adiponectin and lipid traits yielded the greatest evidence for associations for region-based tests.

Informing studies of low-frequency and rare variants

The UK10K-cohorts data allow an empirical evaluation of the relative importance of increasing sample size, genotyping accuracy or variant coverage for increasing power of genetic discoveries across the allele frequency spectrum. In a companion paper⁸ we show that common variants are exhaustively and accurately imputed using current haplotype reference panels, so increasing sample size is likely to be the single most beneficial approach for discovering novel loci driven by common variants. We further show that the UK10K haplotype reference panel, with tenfold more European samples compared to 1000GP, yields substantial improvements in imputation accuracy and coverage for low-frequency and rare variants. To obtain realistic estimates of the power benefit due to imputation with 1000GP+UK10K compared to 1000GP alone, we averaged the smallest value of Beta (the magnitude of a per-allele effect measured in standard deviations) detectable at 80% power, across variants imputable from both reference panels on chromosome 20. Fig. 4a shows sizable reductions in the magnitude of the effect sizes that can be identified at any sample size through use of the UK10K reference panel, compared to the 1000GP panel alone. For instance, for a variant of $\text{MAF} = 0.3\%$ we have equivalent power when imputing from UK10K+1000GP into a 3,621 sample as we have when using the 1000GP imputation panel alone with 10,000 samples.

Similar, although weaker, increases in power were seen for region-based tests of rare variants. Using the WGS autosome data from UK10K, we used simulation to introduce genotype errors into 220 randomly selected regions of 30 variants each. For each variant, errors were simulated to match the MAF and the observed r^2 values between imputation and sequencing, and between whole-exome and whole-genome sequencing (Supplementary Table 11). We modified the SKAT power calculator³⁵ to estimate power both for the true genotypes in a region and the data containing error, and averaged results across the 220 regions (see Supplementary Information). Although absolute power in Fig. 4b is generally poor, we can also see demonstrable power improvements when data are better imputed or are directly sequenced (Fig. 4c).

Tests involving non-coding rare variants may further benefit from aggregation strategies driven by biological annotation that takes into consideration the context- and trait-specific impact of non-coding variation^{36–38}. Exploiting the denser sequence ascertainment of the UK10K-cohorts, we developed a robust approach to quantify fold-enrichment statistics for different categories of non-coding variants compared to null sets matched for minor allele frequency, local linkage disequilibrium and gene density (Supplementary Information). We used this approach to assess the relative contribution of low-frequency and common variants to associations with five exemplar lipid measures (the study did not have sufficient signal for rarer variants). We considered twelve different functional annotation domains, five in or near protein-coding regions and seven main chromatin segmentation states, defined using data from a cell line informative for lipid traits (HepG2; Supplementary Table 12). Low-frequency variants

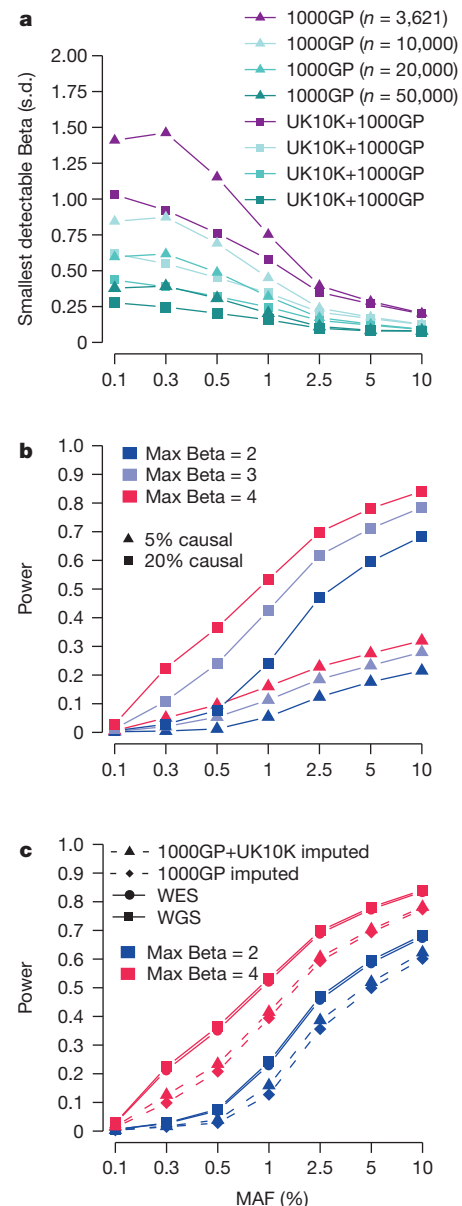


Figure 4 | Power for single-variant and region-based tests. **a**, Strength of single-variant associations detectable at 80% power as a function of MAF and sample size. Using data from chromosome 20⁸, we calculated the smallest value of the strength of association Beta (measured in standard deviations), that would be detectable under a linear dosage model, given the MAF and r^2 of each variant imputable from both the 1000GP and the UK10K+1000GP reference panels, for various sample sizes, n . The averages of these minimum detectable beta values by MAF and sample size are shown. **b**, Power of region-based tests in the UK10K-cohorts sample. Evaluations assume $n = 3,621$, $\alpha = 6.7 \times 10^{-8}$ and that the proportion of causal variants in the regions is either 5% or 20%, for maximum association (Max Beta) in a region = 2, 3, 4 s.d. **c**, Power of region-based tests and the impact of genotype imputation. Ten regions of 30 variants were randomly sampled from each autosome, and then genotype errors were randomly added to the data following observed r^2 values between genotypes from data imputed from different sources (WGS, high depth WES, GWAS imputed against 1000GP, GWAS imputed against the combined reference panel of 1000GP and UK10K; Supplementary Table 11), and matching the MAF of each variant using the same parameters as in **b**, with the proportion of causal variants in the regions set to 20%.

in exonic regions displayed the strongest degree of enrichment (25-fold, compared to fivefold for common variants, Fig. 5), compatible with the effect of purifying selection³⁹. Importantly, however, we showed nearly as strong levels of functional enrichment at both sets of variants for

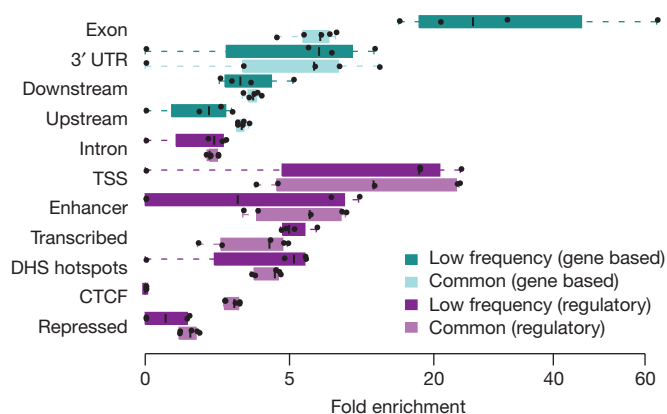


Figure 5 | Enrichment of single-marker associations by functional annotation in the UK10K-cohorts study. Distribution of fold enrichment statistics for single-variant associations of low-frequency (MAF 1–5%) and common (MAF \geq 5%) SNVs in near-genic elements or selected chromatin states and DNase I hotspots (DHS). Boxplots represent distributions of fold enrichment statistics estimated across the five (out of 31 core) traits where at least 10 independent SNVs were associated with the trait at 10^{-7} *P* value (permutation test) threshold (HDL, LDL, TC, APOA1 and APOB). Chromatin state and DHS regions were inferred from ENCODE data in a liver cell line, HepG2, which is informative for lipids. Promoter and 5' UTR are not shown, but corresponding statistics are given in Supplementary Table 12.

several non-coding domains (~10- to 20-fold for transcription start sites, DNase I hotspots and 3' UTRs of genes), confirming the important contribution of non-coding low-frequency alleles to phenotypic trait variance.

Findings from the exome arm of UK10K

In the UK10K-exomes arm studies (see Supplementary Table 13), 5,182 individuals passed sequencing quality control with an average read depth of 80× in the bait regions. We analysed variation discovered in 3,463 disease-affected, unrelated, European-ancestry samples (Supplementary Information). We discovered 842,646 SNVs (of which 1.6% were multiallelic) and 6,067 INDELs. Both variant types were dominated by very rare variants, with more than 60% observed in only one individual. (Extended Data Fig. 6). When compared to European-American samples from the NHLBI Exome Sequencing Project (ESP)³⁹, we found near-complete overlap at sites with MAF \geq 1%: 99% of SNVs that are well covered by both projects and pass quality control are present in both data sets. By contrast, 72% of well-covered SNVs seen only once or twice in UK10K are present in ESP. To inform the functional annotation of these variants, we used the Illumina Body Map to determine if the frequency of LoF and functional variants changed when transcripts are selected based on their expression level (Extended Data Fig. 7). When only consequences from highly expressed transcripts and especially those highly expressed in all the Body Map tissues were considered, LoF and functional changes declined. This demonstrates that the choice of transcript can affect the consequence and this should be taken into account when annotating patient exomes.

The rare disease collection studied 1,000 exomes, or ~125 from each of eight rare diseases. Thus far, 25 novel genetic causes have been identified for five of the eight diseases: ciliopathies ($n = 14$), neuromuscular disorders ($n = 7$), eye malformations ($n = 2$), congenital heart defects ($n = 1$) and intellectual disability ($n = 1$; Supplementary Table 14). Notably, there was marked variation in our ability to identify causal variants based on familial recurrence risk, with the primary factors appearing to be: (1) the proportion of patients with a monogenic cause, (2) the strength of prior information about the mode of inheritance (for example, dominant, recessive), and (3) the

extent of prior knowledge of the relevant functional pathways. In contrast with our success identifying single-diagnostic variants in these rare diseases, our analysis of three complex diseases (obesity, autism spectrum disorder and schizophrenia) on their own did not yield replicating disease-associated loci. This is perhaps unsurprising given expected locus and allelic heterogeneity, and modest sample size⁴⁰. We therefore engaged in a collaborative meta-analysis as part of the Autism Sequencing Consortium⁴¹ which identified 13 associated genes (FDR < 0.01), many of which have been previously shown to cause intellectual disability or developmental disorders. This suggests that rare variation in single genes can have a large role causing a subset of autism spectrum disorder, but these effects only become apparent when large numbers of individuals are studied.

We also used the UK10K-exomes sequence data to explore the occurrence of incidental findings. We focused on disease-specific genes identified in current guidelines for the analysis of exome/whole-genome data by the American College of Medical Genetics and Genomics (ACMG)⁴², and used objective criteria described in the Supplementary Information. We identified a total of 29 distinct reportable variants affecting a total of 2.3% of the UK10K cases considered in this analysis (42 out of 1,805 individuals), a number similar to previous estimates (2% estimate in adults of European ancestry⁴³). The incidental findings were predominantly associated with cardiovascular disorders (Supplementary Table 15).

Two main challenges of reporting incidental findings from whole-exome surveys emerge. The need for clinical expertise, the difficulty of interpreting a fraction of variants, and the lack of completeness of the ClinVar database⁴⁴ all highlighted the need to further consolidate knowledge from the community into freely accessible and more exhaustive databases. Furthermore, for some disorders, the frequency of carriers is likely to be too high compared to the disease frequency, despite our strict assessment criteria. This suggests that reported estimates of the penetrance of recognized variants for specific disorders are too high. Given these challenges, we suggest that, in the absence of additional evidence, scientific publications describing proposed penetrant associations for rare variants need to be complemented by accurate estimates of population frequencies.

Conclusions

In summary we have generated a high-quality whole-genome sequence data repository including 24 million novel variants from nearly 4,000 European-ancestry individuals. We showed that the UK10K haplotype reference panel greatly increases accuracy and coverage of low-frequency and rare variants compared to existing panels such as the 1000GP phase 1 panel. We carried out a large-scale empirical exploration of association testing of common, low-frequency and rare genetic variants with a large variety of biomedically important quantitative traits. For each of the different association scenarios tested, we report first examples of novel alleles associated with lipid and adiponectin traits. This provides proof-of-principle evidence on the value of the large-scale sequencing data for complex traits, while also indicating that there are few low-frequency large effect 'quick wins' that make substantial contributions to population trait variation and that can be discovered from sequencing studies of few thousands individuals. Our power calculations, informed by the sequence data, provide realistic estimates of the benefit of sequencing versus imputation in future association studies. Finally, rare variation tests showed limited evidence for confounding owing to population stratification at the traits investigated, likely to be due to a weakening of historical patterns of population structure in the current general UK population⁴⁵.

Overall, this effort has given us both new genomic tools¹² and insights into the role of low-frequency and rare variation on human complex traits, and will inform strategies for future association studies. Our exploration of non-coding variants supports the need for incorporating functional genome information in association tests of rare

variants outside protein-coding regions. Improved study power through larger numbers, and a better understanding of the observed heterogeneity in allelic architecture between different loci, are likely to provide the best route forward to describe the contribution of rare variants to phenotypic variance in health and disease, and for assessing their utility in healthcare.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 March 2015; accepted 17 July 2015.

Published online 14 September; corrected online 30 September 2015 (see full-text HTML version for details).

- Manolio, T. A. Bringing genome-wide association findings into clinical use. *Nature Rev. Genet.* **14**, 549–558 (2013).
- Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8**, e1002793 (2012).
- Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
- Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genet.* **46**, 220–224 (2014).
- 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- Lange, L. A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94**, 233–245 (2014).
- Hindorf, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
- Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Commun.* **6**, 8111 (2015).
- Zheng, H. *et al.* Whole-genome sequencing identifies *EN1* as a determinant of bone density and fracture. *Nature* <http://dx.doi.org/10.1038/nature14878> (2015).
- Taylor, P. N. *et al.* Whole-genome sequence-based analysis of thyroid function. *Nature Commun.* **6**, 5681 (2015).
- Timpson, N. J. *et al.* A rare variant in *APOC3* is associated with plasma triglyceride and VLDL levels in Europeans. *Nature Commun.* **5**, 4871 (2014).
- Geijs, M. *et al.* An interactive genome browser of association results from the UK10K cohorts project. *Bioinformatics*. <http://dx.doi.org/10.1093/bioinformatics/btv491> (2015).
- Boyd, A. *et al.* Cohort Profile: the ‘children of the 90s’—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **42**, 111–127 (2013).
- Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **16**, 144–149 (2013).
- Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21**, 940–951 (2011).
- Wheeler, E. *et al.* Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nature Genet.* **45**, 513–517 (2013).
- Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genet.* **47**, 435–444 (2015).
- Williams, F. M. *et al.* Genes contributing to pain sensitivity in the normal population: an exome sequencing study. *PLoS Genet.* **8**, e1003095 (2012).
- Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genet.* **46**, 818–825 (2014).
- Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
- Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* **38**, 281–290 (2014).
- Jørgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjaerg-Hansen, A. Loss-of-function mutations in *APOC3* and risk of ischemic vascular disease. *N. Engl. J. Med.* **371**, 32–41 (2014).
- The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute. Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. *N. Engl. J. Med.* **371**, 22–31 (2014).
- McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
- Park, J. H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl Acad. Sci. USA* **108**, 18026–18031 (2011).
- Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42**, 565–569 (2010).
- Nordestgaard, B. G., Benn, M., Schnohr, P. & Tybjaerg-Hansen, A. Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women. *J. Am. Med. Assoc.* **298**, 299–308 (2007).
- Whittall, R. A., Matheus, S., Cranston, T., Miller, G. J. & Humphries, S. E. The intron 14 2140+5G>A variant in the low density lipoprotein receptor gene has no effect on plasma cholesterol levels. *J. Med. Genet.* **39**, e57 (2002).
- Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
- Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* **44**, 293–308 (2010).
- Wu, C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
- Liu, D. J. & Leal, S. M. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *Am. J. Hum. Genet.* **91**, 585–596 (2012).
- Morisaki, H. *et al.* *CDH13* gene coding T-cadherin influences variations in plasma adiponectin levels in the Japanese population. *Hum. Mutat.* **33**, 402–410 (2012).
- Dastani, Z. *et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, e1002607 (2012).
- Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **488**, 57–74 (2012).
- Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnol.* **30**, 224–226 (2012).
- Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
- Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).
- Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013).
- Kaye, J. *et al.* Managing clinically significant findings in research: the UK10K example. *Eur. J. Hum. Genet.* **22**, 1100–1104 (2014).
- Amendola, L. M. *et al.* Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res.* **25**, 305–315 (2015).
- Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
- Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
- Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genet.* **44**, 243–246 (2012).
- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
- Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
- Benjamini, Y. & Hochberg, Y. controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

Supplementary Information is available in the online version of the paper.

Acknowledgements This study makes use of data generated by the UK10K Consortium. The Wellcome Trust provided funding for UK10K (WT091310). Additional grant support and acknowledgements can be found in the Supplementary Information.

Author Contributions Project management: D.M., K.R.; designed individual studies and contributed data: A.A., A.Do., A.G.M., A.I., A.Ma., A.McI., A.McQ., A.Mor., A.O., A.R.F., A.T.H., A.Val., A.Var., B.H.S., B.N., C.B., C.C., C.M.v., C.W., C.L., D.A., D.B., D.B.S., D.Co., D.Cu., D.Ge., D.Gr., D.H., D.J.P., D.R.F., D.S.-C., D.S., D.T., E.M.v., E.St., E.Z., F.M., F.Z., G.B., G.C.I., G.D., G.G., G.L., G.Mal., G.S., G.Z., H.Gu., H.M.M., H.W., I.L., I.N.M.D., I.S.F., J.B., J.C., J.C.C., J.H., J.J., J.Keo., J.L.M., J.Lö., J.Lu., J.Mo., J.R.P., J.S.K., J.Suv., J.Wal., K.A.W., K.Ch., K.J.W., K.N., L.G., F.L.R., L.S., M.A., M.Be., M.C.O., M.Ca., M.Co., M.D.T., M.E.K., M.J.O., M.M., M.S., M.T., N.C., N.J.T., N.R., N.Sc., N.So., O.S., P.Be., P.Bo., P.G., P.Ho., P.M., P.Sc., P.W., R.A., R.B., R.K.S., R.M., R.S., S.Bh., S.Ci., S.Cu., S.E.H., S.G.W., S.I.S., S.O., S.R., T.D.S., T.G., T.P., T.W., V.I., V.Pa., W.M., UCLEB Consortium†; generated and/or quality controlled sequence data: A.Da., A.K.-K., C.J., Co.L., D.K.J., D.M., F.Z., G.Co., G.W., H.L., J.H., J.Li., J.Mas., J.St., J.Sun., J.T., K.Wo., M.A.Q., P.C., P.D., P.E., P.F., R.D., Ru.L., Ry.L., S.Ba., S.E., S.McC., T.C., T.K., Xi.G., Y.D.; designed new statistical or bioinformatics tools: A.C., A.H., B.H., C.M.T.G., C.X., E.Bi., E.Z., G.R.S.R., H.S., I.D., I.T., J.Mar., K.O., N.So., Ru.L., S.Me., T.D., T.H., V.I.; analysed the data and provided critical interpretation of results: A.D.-W., A.H., A.M.V., A.Moa., A.P., A.S., J.B.R., B.S., C.A., C.M.T.G., C.K.R., C.S.F., C.W., D.E., D.G.M., D.L., E.Bo., E.Se., E.W., E.Z., F.K., F.P., F.Z., G.Mar., G.R.S.R., H.Z., I.B., I.M., I.T., J.C.B., J.F., J.H., J.Kem., J.L.M., J.Mo., J.R.B.P., J.Y., K.Ca., K.P., K.S., K.Wa., K.Wo., L.Ch., L.Cr., L.P., L.Q., L.R.L., L.S., L.V.W., M.Co., M.E.H., M.F., M.G., M.Le., M.S.-A., M.S., N.J.T., N.M., N.S.O., O.P., P.D., P.Hy., P.M.V., P.Sy., P.V., R.C., R.C.P., R.D., R.E., R.L.R., R.T., R.S., S.-Y.S., S.A., S.E.H., S.G.W., S.McC., S.Me., S.P., S.S., T.G., V.I., V.P.I., Y.J., Y.M.; ethics: A.K., C.S., D.M., D.R.F., F.M., H.Gr., J.Ka., K.K., F.L.R., M.Bo., M.E.H., N.J.T., P.Bo., R.D., R.K.S., T.D.S.; designed and/or managed the study: A.P., J.B.R., Co.L., D.M., D.R.F., E.Z., G.D.-S., I.S.F., J.C.B., J.Ka., J.St., K.K., M.E.H., M.J.O., N.J.T., N.So., R.D., S.McC., T.D.S.; wrote the manuscript: A.H., J.B.R., C.M.T.G., C.X., D.L., E.Z., I.B., J.C.B., J.F., J.H., J.L.M., J.R.B.P., K.Wa., L.Cr., M.E.H., M.F., N.J.T., N.So., P.D., R.D., Ru.L., S.E.H., S.McC., S.S., V.I., V.P.I., Y.M.

Author Information Data access form is available at http://www.uk10k.org/data_access.html, raw and processed data files at <https://www.ebi.ac.uk/ega/>, imputation panel at <https://www.ebi.ac.uk/ega/>, UK10K Genome Browser at <http://www.uk10k.org/dalliance.html>, single-marker loci navigator at http://fathmm.biocompute.org.uk/UK10K_Browser/ and dynamic power calculator at http://fathmm.biocompute.org.uk/UK10K_Browser/Power.htm. All sequence and

phenotype data were deposited to the European Genome-Phenome archive (EGA, <https://www.ebi.ac.uk/ega/>), with accession numbers EGAD00001000740, EGAD00001000789, EGAD00001000741, EGAD00001000790, EGAD00001000776, EGAD00001000433, EGAD00001000434, EGAD00001000435, EGAD00001000436, EGAD00001000613, EGAD00001000614, EGAD00001000437, EGAD00001000438, EGAD00001000615, EGAD00001000439, EGAD00001000440, EGAD00001000441, EGAD00001000442, EGAD00001000443, EGAD00001000430, EGAD00001000431, EGAD00001000432, EGAD00001000429, EGAD00001000413, EGAD00001000414, EGAD00001000415, EGAD00001000416, EGAD00001000417, EGAD00001000418, EGAD00001000419 and EGAD00001000420. A breakdown of studies is given in Supplementary Table 13. All study participants provided informed consent. Details of REC approvals are given in Supplementary Table 17.

Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.S. (ns6@sanger.ac.uk) or R.D. (rd@sanger.ac.uk).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

The UK10K Project Consortium

Writing group

Klaudia Walter^{1*}, Josine L. Min^{3*}, Jie Huang^{1*}, Lucy Crooks^{1,4*}, Yasin Memari¹, Shane McCarthy¹, John R. B. Perry^{5,6}, ChangJiang Xu^{7,8}, Marta Futema⁹, Daniel Lawson¹⁰, Valentina Iotchkova^{1,11}, Stephan Schiffels¹, Audrey E. Hendricks^{1,12}, Petr Danecek¹, Rui Li^{7,13,14}, James Floyd^{1,15}, Louise V. Wain¹⁶, Inês Barroso^{1,17}, Steve E. Humphries¹, Matthew E. Hurles¹, Eleftheria Zeggini¹, Jeffrey C. Barrett¹, Vincent Plagnol^{1,8}, J. Brent Richards^{5,7,8,13,14}, Celia M. T. Greenwood^{7,8,14,19}, Nicholas J. Timpson³, Richard Durbin¹, Nicole Soranzo^{1,2}

Production group

Senduran Bala¹, Peter Clapham¹, Guy Coates¹, Tony Cox¹, Allan Daly¹, Petr Danecek¹, Yuanping Du²⁰, Richard Durbin¹, Sarah Edkins¹, Peter Ellis¹, Paul Ficek^{1,11}, Xiaosen Guo^{20,21}, Xueqin Guo²⁰, Liren Huang²⁰, David K. Jackson¹, Chris Joyce¹, Thomas Keane¹, Anja Kolb-Kococinski¹, Cordelia Langford¹, Yingrui Li²⁰, Jieqin Liang²⁰, Hong Lin²⁰, Ryan Liu²², John Maslen¹, Shane McCarthy (co-chair)¹, Dawn Muddyman¹, Michael A. Quail¹, Jim Stalker (co-chair)¹, Jianping Sun^{7,8}, Jing Tian²⁰, Guangbiao Wang²⁰, Jun Wang^{20,21,23,24,25}, Yu Wang²⁰, Kim Wong¹, Pingbo Zhang²⁰

Cohorts group

Inês Barroso^{1,17}, Ewan Birney¹¹, Chris Boustred²⁶, Lu Chen^{1,2}, Gail Clement⁵, Massimiliano Cocca^{27,28}, Petr Danecek¹, George Davey Smith³, Ian N. M. Day²⁹, Aaron Day-Williams^{1,30}, Thomas Down^{1,31}, Ian Dunham¹¹, Richard Durbin¹, David M. Evans^{3,32}, Tom R. Gaunt³, Matthias Geihs¹, Celia M. T. Greenwood^{7,8,14,19}, Deborah Hart⁵, Audrey E. Hendricks^{1,12}, Bryan Howie³³, Jie Huang¹, Tim Hubbard^{1,31}, Pirro Hysi¹, Valentina Iotchkova^{1,11}, Yalda Jamshidi³⁴, Konrad J. Karczewski^{35,36}, John P. Kemp^{3,32}, Genevieve Lachance⁵, Daniel Lawson¹⁰, Monkol Lek³⁵, Margarida Lopes^{1,37,38}, Daniel G. MacArthur^{35,36}, Jonathan Marchini^{37,39}, Massimo Mangino^{5,40}, Iain Mathieson⁴¹, Shane McCarthy¹, Yasin Memari¹, Sarah Metrustry⁵, Josine L. Min³, Alireza Moayeri^{5,42}, Dawn Muddyman¹, Kate Northstone³, Kalliope Panoutsopoulou¹, Lavinia Paternoster³, John R. B. Perry^{5,6}, Lydia Quayle⁵, J. Brent Richards (co-chair)^{5,7,8,13,14}, Susan Ring⁴³, Graham R. S. Ritchie^{1,11}, Stephan Schiffels¹, Hashem A. Shihab³, So-Youn Shin^{1,3}, Kerrin S. Small⁵, Maria Soler Artigas¹⁶, Nicole Soranzo (co-chair)^{1,2}, Lorraine Southam^{1,37}, Timothy D. Spector⁵, Beate St Pourcain^{3,44,45}, Gabriela Surdulescu⁵, Ioanna Tachmazidou¹, Nicholas J. Timpson (co-chair)³, Martin D. Tobin^{16,46}, Ana M. Valdes⁵, Peter M. Visscher^{32,47}, Louise V. Wain¹⁶, Klaudia Walter¹, Kirsten Ward⁵, Scott G. Wilson^{5,48,49}, Kim Wong¹, Jian Yang^{32,47}, Eleftheria Zeggini¹, Feng Zhang⁵, Hou-Feng Zheng^{7,13,14}

Neurodevelopmental disorders group

Richard Anney⁵⁰, Muhammad Ayub⁵¹, Jeffrey C. Barrett¹, Douglas Blackwood⁵², Patrick F. Bolton^{53,54,55}, Gerome Breen^{54,55}, David A. Collier^{55,56}, Nick Craddock⁵⁷, Lucy Crooks^{1,4}, Sarah Curran^{53,58,59}, David Curtis⁶⁰, Richard Durbin¹, Louise Gallagher⁵⁰, Daniel Geschwind⁶¹, Hugh Gurling^{62,63}, Peter Holmans⁵⁷, Irene Lee⁶³, Jouko Lönnqvist⁶⁴, Shane McCarthy¹, Peter McGuffin⁵⁵, Andrew M. McIntosh⁵², Andrew G. McKechanie^{52,65}, Andrew McQuillin⁶², James Morris¹, Dawn Muddyman¹, Michael C. O'Donovan⁵⁷, Michael J. Owen (co-chair)⁵⁷, Aarno Palotie (co-chair)^{1,66,67}, Jeremy R. Parr⁶⁸, Tiina Paunio^{64,69}, Olli Pietiläinen^{1,64,66}, Karola Rehnström¹, Sally I. Sharp⁶², David Skuse⁶³, David St Clair⁷⁰, Jaana Suvisaari⁶⁴, James T. R. Walters⁵⁷, Hywel J. Williams^{57,71}

Obesity group

Inês Barroso (co-chair)^{1,17}, Elena Bochukova¹⁷, Rebecca Bounds¹⁷, Anna Dominiczak⁷², Richard Durbin¹, I. Sadaf Farooqi (co-chair)¹⁷, Audrey E. Hendricks^{1,12},

Julia Keogh¹⁷, Gaëlle Marenne¹, Shane McCarthy¹, Andrew Morris⁷³, Dawn Muddyman¹, Stephen O'Rahilly¹⁷, David J. Porteous⁷⁴, Blair H. Smith⁷⁵, Ioanna Tachmazidou¹, Eleanor Wheeler¹, Eleftheria Zeggini¹

Rare disease group

Saeed Al Turki^{1,76}, Carl A. Anderson¹, Dinu Antony⁷⁷, Inês Barroso^{1,17}, Phil Beales⁷⁷, Jamie Bentham⁷⁸, Shoumo Bhattacharya⁷⁸, Mattia Calissano⁷⁹, Keren Carss¹, Krishna Chatterjee¹⁷, Sebhattin Cirak^{79,80}, Catherine Cosgrove⁷⁸, Richard Durbin¹, David R. Fitzpatrick (co-chair)⁸¹, James Floyd^{1,15}, A. Reghan Foley⁷⁹, Christopher S. Franklin¹, Marta Futema⁹, Detelina Grozeva⁸², Steve E. Humphries⁹, Matthew E. Hurles (co-chair)¹, Shane McCarthy¹, Hannah M. Mitchison⁷⁷, Dawn Muddyman¹, Francesco Muntoni⁷⁹, Stephen O'Rahilly¹⁷, Alexandros Onoufriadi³¹, Victoria Parker¹⁷, Felicity Payne¹, Vincent Plagnol¹⁸, F. Lucy Raymond⁸², Nicola Roberts⁸², David B. Savage¹⁷, Peter Scambler⁷⁷, Miriam Schmidts^{77,83}, Nadia Schoenmakers¹⁷, Robert K. Semple¹⁷, Eva Serra¹, Olivera Spasic-Boskovic⁸², Elizabeth Stevens⁷⁹, Margriet van Kogelenberg¹, Parthiban Vijayarangakannan¹, Klaudia Walter¹, Kathleen A. Williamson⁸¹, Crispian Wilson⁸², Tamieka Whyte⁷⁹

Statistics group

Antonio Ciampi⁸, Celia M. T. Greenwood (co-chair)^{7,8,14,19}, Audrey E. Hendricks^{1,12}, Rui Li^{7,13,14}, Sarah Metrustry⁵, Karim Ouakacha⁸⁴, Ioanna Tachmazidou¹, ChangJiang Xu^{7,8}, Eleftheria Zeggini (co-chair)¹

Ethics group

Martin Bobrow⁸², Patrick F. Bolton^{53,54,55}, Richard Durbin¹, David R. Fitzpatrick⁸¹, Heather Griffin⁸⁵, Matthew E. Hurles (co-chair)¹, Jane Kaye (co-chair)⁸⁵, Karen Kennedy^{1,86}, Alastair Kent⁸⁷, Dawn Muddyman¹, Francesco Muntoni⁷⁹, F. Lucy Raymond⁸², Robert K. Semple¹⁷, Carol Smee¹, Timothy D. Spector⁵, Nicholas J. Timpson³

Incidental findings group

Ruth Charlton⁸⁸, Rosemary Ekong⁸⁹, Marta Futema⁹, Steve E. Humphries⁹, Farrah Khawaja⁹⁰, Luis R. Lopes^{91,92}, Nicola Migone⁹³, Stewart J. Payne⁹⁴, Vincent Plagnol (chair)¹⁸, Rebecca C. Pollitt⁹⁵, Sue Povey⁸⁹, Cheryl K. Ridout⁹⁶, Rachel L. Robinson⁸⁸, Richard H. Scott^{77,97}, Adam Shaw⁹⁸, Petros Syrris⁹¹, Rohan Taylor⁹⁰, Anthony M. Vandersteern⁹⁹

Management committee

Jeffrey C. Barrett¹, Inês Barroso^{1,17}, George Davey Smith³, Richard Durbin (chair)¹, I. Sadaf Farooqi¹⁷, David R. Fitzpatrick⁸¹, Matthew E. Hurles¹, Jane Kaye⁸⁵, Karen Kennedy^{1,86}, Cordelia Langford¹, Shane McCarthy¹, Dawn Muddyman¹, Michael J. Owen⁵⁷, Aarno Palotie^{1,66,67}, J. Brent Richards^{5,7,8,13,14}, Nicole Soranzo^{1,2}, Timothy D. Spector⁵, Jim Stalker¹, Nicholas J. Timpson³, Eleftheria Zeggini¹

Lipid meta-analysis group

Antoinette Amuzu¹⁰⁰, Juan Pablo Casas^{91,100}, John C. Chambers¹⁰¹, Massimiliano Cocca^{27,28}, George Dedoussis¹⁰², Giovanni Gambaro¹⁰³, Paolo Gasparini^{27,28,104}, Tom R. Gaunt³, Jie Huang¹, Valentina Iotchkova^{1,11}, Aaron Isaacs¹⁰⁵, Jon Johnson¹⁰⁶, Marcus E. Kleber¹⁰⁷, Jaspal S. Kooner¹⁰⁸, Claudia Langenberg¹⁰⁹, Jian'an Luan¹⁰⁹, Giovanni Malarba¹¹⁰, Winfried März^{111,112,113}, Angela Matchan¹, Josine L. Min³, Richard Morris¹¹⁴, Børge G. Nordestgaard^{115,116}, Marianne Benn^{115,116}, Susan Ring⁴³, Robert A. Scott¹⁰⁹, Nicole Soranzo^{1,2}, Lorraine Southam^{1,37}, Nicholas J. Timpson³, The UCLEB Consortium¹, Daniela Toniolo¹¹⁷, Michela Traglia¹¹⁷, Anne Tybjaerg-Hansen^{116,118}, Cornelia M. van Duijn¹⁰⁵, Elisabeth M. van Leeuwen¹⁰⁵, Anette Varbo^{115,116}, Peter Whincup¹¹⁹, Gianluigi Zaza¹²⁰, Eleftheria Zeggini¹, Weihua Zhang¹⁰¹

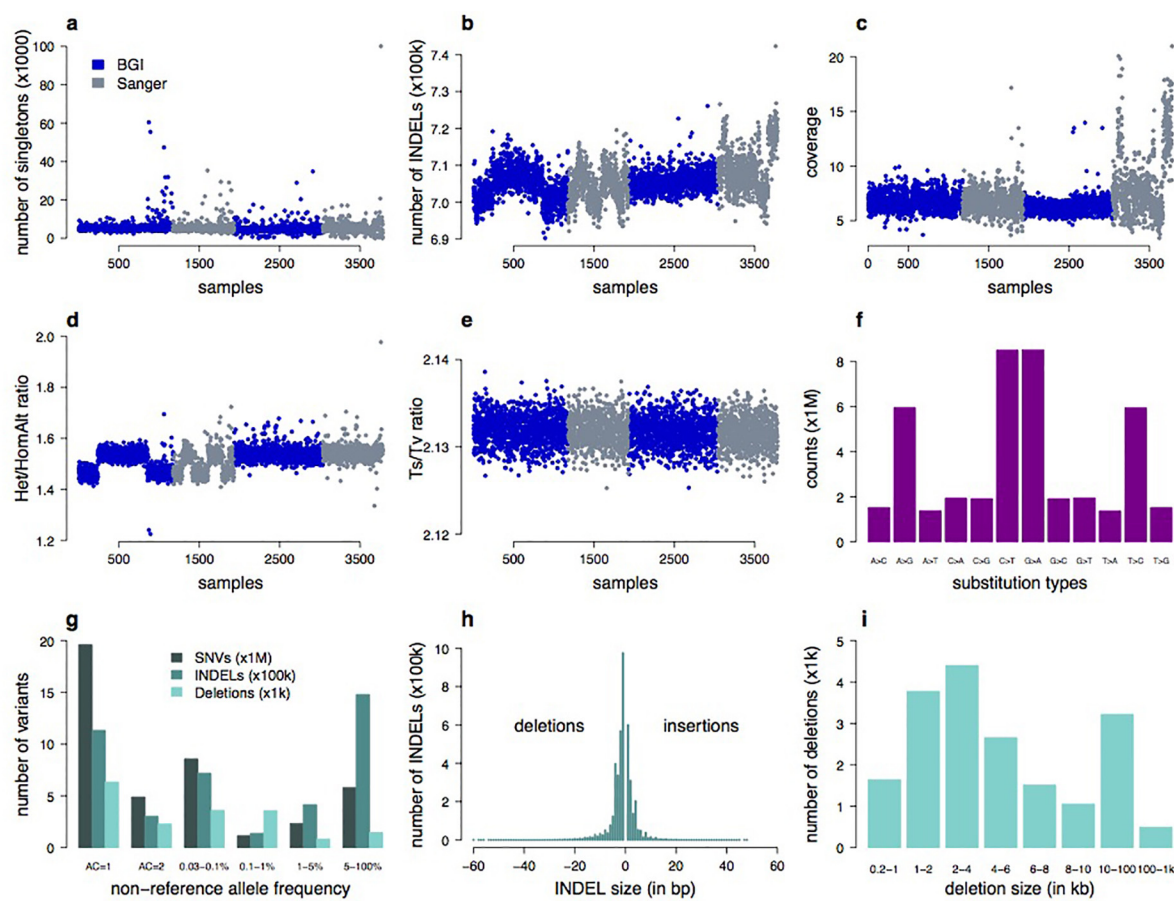
¹The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1HH, Cambridge, UK. ²Department of Haematology, University of Cambridge, Long Road, Cambridge CB2 0PT, UK. ³MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. ⁴Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Foundation Trust, Western Bank, Sheffield S10 2TH, UK. ⁵The Department of Twin Research & Genetic Epidemiology, King's College London, St Thomas' Campus, Lambeth Palace Road, London SE1 7EH, UK. ⁶MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Box 285, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. ⁷Lady Davis Institute, Jewish General Hospital, Montreal, Quebec H3T 1E2, Canada. ⁸Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec H3A 1A2, Canada. ⁹Cardiovascular Genetics, BHF Laboratories, Rayne Building, Institute of Cardiovascular Sciences, University College London, London WC1E 6JJ, UK. ¹⁰Schools of Mathematics and Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. ¹¹European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ¹²Department of Mathematical and Statistical Sciences, University of Colorado, Denver, Colorado 80204, USA. ¹³Department of Medicine, Jewish General Hospital, McGill University, Montreal, Quebec H3A 1B1, Canada. ¹⁴Department of Human Genetics, McGill University, Montreal, Quebec H3A 1B1, Canada. ¹⁵The Genome Centre, John Vane Science Centre, Queen Mary, University of London, Charterhouse Square, London EC1M 6BQ, UK. ¹⁶Departments of Health Sciences and Genetics, University of Leicester, Leicester LE1 7RH, UK. ¹⁷University of Cambridge Metabolic Research Laboratories, and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. ¹⁸University College London (UCL) Genetics Institute (UGI) Gower Street, London WC1E 6BT, UK. ¹⁹Department of Oncology, McGill University, Montreal, Quebec H2W 1S6, Canada. ²⁰BGI-Shenzhen, Shenzhen 518083, China. ²¹Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark. ²²BGI-Europe, London

EC2M 4YE, UK. ²³Princess Al Jawhara Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, P.O. Box 80200, Jeddah 21589, Saudi Arabia. ²⁴Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. ²⁵Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong. ²⁶North East Thames Regional Genetics Service, Great Ormond Street Hospital NHS Foundation Trust, London WC1N 3JH, UK. ²⁷Medical Genetics, Institute for Maternal and Child Health IRCCS "Burlo Garofolo", 34100 Trieste, Italy. ²⁸Department of Medical, Surgical and Health Sciences, University of Trieste, 34100 Trieste, Italy. ²⁹Bristol Genetic Epidemiology Laboratories, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. ³⁰Computational Biology & Genomics, Biogen Idec, 14 Cambridge Center, Cambridge, Massachusetts 02142, USA. ³¹Department of Medical and Molecular Genetics, Division of Genetics and Molecular Medicine, King's College London School of Medicine, Guy's Hospital, London SE1 9RT, UK. ³²University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland 4102, Australia. ³³Adaptive Biotechnologies Corporation, Seattle, Washington 98102, USA. ³⁴Human Genetics Research Centre, St George's University of London, London SW17 0RE, UK. ³⁵Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ³⁶Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. ³⁷Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. ³⁸Illumina Cambridge Ltd, Chesterford Research Park, Cambridge CB10 1XL, UK. ³⁹Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. ⁴⁰National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' Foundation Trust, London SE1 9RT, UK. ⁴¹Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴²Institute of Health Informatics, Farr Institute of Health Informatics Research, University College London (UCL), 222 Euston Road, London NW1 2DA, UK. ⁴³ALSPAC & School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol BS8 2BN, UK. ⁴⁴School of Oral and Dental Sciences, University of Bristol, Lower Maudlin Street, Bristol BS1 2LY, UK. ⁴⁵School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK. ⁴⁶National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester LE3 9QP, UK. ⁴⁷Queensland Brain Institute, University of Queensland, Brisbane, Queensland 4072, Australia. ⁴⁸School of Medicine and Pharmacology, University of Western Australia, Perth, Western Australia 6009, Australia. ⁴⁹Department of Endocrinology and Diabetes, Sir Charles Gairdner Hospital, Nedlands, Western Australia 6009, Australia. ⁵⁰Department of Psychiatry, Trinity Centre for Health Sciences, St James Hospital, James's Street, Dublin 8, Ireland. ⁵¹Division of Developmental Disabilities, Department of Psychiatry, Queen's University, Kingston, Ontario N6C 0A7, Canada. ⁵²Division of Psychiatry, The University of Edinburgh, Royal Edinburgh Hospital, Edinburgh EH10 5HF, UK. ⁵³Department of Child Psychiatry, Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, London SE5 8AF, UK. ⁵⁴NIHR BRC for Mental Health, Institute of Psychiatry, Psychology and Neuroscience and SLam NHS Trust, King's College London, 16 De Crespigny Park, London SE5 8AF, UK. ⁵⁵MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, Denmark Hill, London SE5 8AF, UK. ⁵⁶Lilly Research Laboratories, Eli Lilly & Co. Ltd., Erl Wood Manor, Sunninghill Road, Windlesham GU20 6PH, UK. ⁵⁷MRC Centre for Neuropsychiatric Genetics & Genomics, Institute of Psychological Medicine & Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff CF24 4HQ, UK. ⁵⁸University of Sussex, Brighton BN1 9RH, UK. ⁵⁹Sussex Partnership NHS Foundation Trust, Swandean, Arundel Road, Worthing BN13 3EP, UK. ⁶⁰University College London (UCL), UCL Genetics Institute, Darwin Building, Gower Street, London WC1E 6BT, UK. ⁶¹UCLA David Geffen School of Medicine, Los Angeles, California 90095, USA. ⁶²University College London (UCL), Molecular Psychiatry Laboratory, Division of Psychiatry, Gower Street, London WC1E 6BT, UK. ⁶³Behavioural and Brain Sciences Unit, UCL Institute of Child Health, London WC1N 1EH, UK. ⁶⁴National Institute for Health and Welfare (THL), Helsinki FI-00271, Finland. ⁶⁵The Patrick Wild Centre, The University of Edinburgh, Edinburgh EH10 5HF, UK. ⁶⁶Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki FI-00014, Finland. ⁶⁷Program in Medical and Population Genetics and Genetic Analysis Platform, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02132, USA. ⁶⁸Institute of Neuroscience, Henry Wellcome Building for Neuroecology, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. ⁶⁹University of Helsinki, Department of Psychiatry, Helsinki FI-00014, Finland. ⁷⁰Institute of Medical Sciences, University of Aberdeen, Aberdeen AB25 2ZD, UK. ⁷¹The Centre for Translational Omics – GOSgene, UCL Institute of Child Health, London WC1N 1EH, UK. ⁷²Institute of Cardiovascular and Medical Sciences, University of Glasgow, Wolfson Medical School Building, University Avenue, Glasgow, G12 8QQ, UK. ⁷³Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, 9 Little France Road, Edinburgh EH16 4UX, UK. ⁷⁴Centre for Genomic and Experimental Medicine, Institute of Genetics and Experimental Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. ⁷⁵Mackenzie Building, Kirsty Semple Way, Ninewells Hospital and Medical School, Dundee DD2 4RB, UK. ⁷⁶Department of Pathology, King Abdulaziz Medical City, P.O. Box 22490, Riyadh 11426, Saudi Arabia. ⁷⁷Genetics and Genomic Medicine and Birth Defects Research Centre, UCL Institute of Child Health, London WC1N 1EH, UK. ⁷⁸Department of Cardiovascular Medicine and Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. ⁷⁹Dubowitz Neuromuscular Centre, UCL Institute of Child Health & Great Ormond Street Hospital, London WC1N 1EH, UK. ⁸⁰Institut für Humangenetik, Uniklinik Köln, Kerpener Strasse 34, 50931 Köln, Germany. ⁸¹MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, at the University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. ⁸²Academic Laboratory of Medical Genetics, Box 238, Lv 6 Addenbrooke's Treatment Centre, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. ⁸³Human Genetics Department, Radboudumc and Radboud Institute for Molecular Life Sciences (RIMLS), Geert Grooteplein 25, Nijmegen 6525 HP, The Netherlands. ⁸⁴Department of Mathematics, Université de Québec à Montréal, Montréal, Québec H3C 3P8, Canada. ⁸⁵HeLEX – Centre for Health, Law and Emerging Technologies, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK. ⁸⁶National Cancer Research Institute, Angel Building, 407 St John Street, London EC1V 4AD, UK. ⁸⁷Genetic Alliance UK, 4D Leroy House, 436 Essex Road, London N1 3QP, UK. ⁸⁸Leeds Genetics Laboratory, St James University Hospital, Beckett Street, Leeds LS9 7TF, UK. ⁸⁹University College London (UCL) Department of Genetics, Evolution & Environment (GEE), Gower Street, London WC1E 6BT, UK. ⁹⁰SW Thames Regional Genetics Lab, St George's University, Cranmer Terrace, London SW17 0RE, UK. ⁹¹Institute of Cardiovascular Science, University College London, Gower Street, London WC1E 6BT, UK. ⁹²Cardiovascular Centre of the University of Lisbon, Faculty of Medicine, University of Lisbon, Avenida Professor Egas Moniz, 1649-028 Lisbon, Portugal. ⁹³Department of Medical Sciences, University of Torino, 10124 Torino, Italy. ⁹⁴North West Thames Regional Genetics Service, Kennedy-Galton Centre, Northwick Park Hospital, Watford Road, Harrow HA1 3UJ, UK. ⁹⁵Connective Tissue Disorders Service, Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Foundation Trust, Western Bank, Sheffield S10 2TH, UK. ⁹⁶Molecular Genetics, Viapath at Guy's Hospital, London SE1 9RT, UK. ⁹⁷Department of Clinical Genetics, Great Ormond Street Hospital, London, WC1N 3JH, UK. ⁹⁸Clinical Genetics, Guy's & St Thomas' NHS Foundation Trust, London SE1 9RT, UK. ⁹⁹Maritime Medical Genetics Service, 5850/5980 University Avenue, PO Box 9700, Halifax, Nova Scotia B3K 6R8, Canada. ¹⁰⁰London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. ¹⁰¹The Department of Epidemiology and Biostatistics, Imperial College London, St Mary's campus, Norfolk Place, Paddington, London W2 1PG, UK. ¹⁰²Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens 17671, Greece. ¹⁰³Division of Nephrology and Dialysis, Institute of Internal Medicine, Renal Program, Columbus-Gemelli University Hospital, Catholic University, 00168 Rome, Italy. ¹⁰⁴Experimental Genetics Division, Sidra, P.O. Box 26999 Doha, Qatar. ¹⁰⁵Genetic Epidemiology Unit, Department of Epidemiology, Erasmus MC, Rotterdam 3000 CA, Netherlands. ¹⁰⁶Department of Quantitative Social Science, UCL Institute of Education, University College London, 20 Bedford Way, London WC1H 0AL, UK. ¹⁰⁷Vth Department of Medicine, Medical Faculty, Mannheim 68167, Germany. ¹⁰⁸National Heart and Lung Institute, Imperial College London, London W12 0NN, UK. ¹⁰⁹MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. ¹¹⁰Biology and Genetics, Department of Life and Reproduction Sciences, University of Verona, 37134 Verona, Italy. ¹¹¹Clinical Institute of Medical and Chemical Laboratory Diagnostics, Medical University of Graz, Graz 8036, Austria. ¹¹²Synlab Academy, Synlab Services GmbH, D-68161 Mannheim, Germany. ¹¹³Medical Clinic V (Nephrology, Hypertensiology, Rheumatology, Endocrinology, Diabetology), Mannheim Medical Faculty, Heidelberg University, Mannheim 68167, Germany. ¹¹⁴School of Social and Community Medicine, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK. ¹¹⁵Department of Clinical Biochemistry and The Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark. ¹¹⁶The Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark. ¹¹⁷Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Milan 20132, Italy. ¹¹⁸Department of Clinical Biochemistry KB3011, Rigshospitalet, Copenhagen University Hospital, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. ¹¹⁹Population Health Research Institute, St George's University of London, London SW17 0RE, UK. ¹²⁰Renal Unit, Department of Medicine, University of Verona, 37126 Verona, Italy.

†A list of authors and affiliations appears in the Supplementary Information.

‡Deceased.

*These authors contributed equally to this work.



j

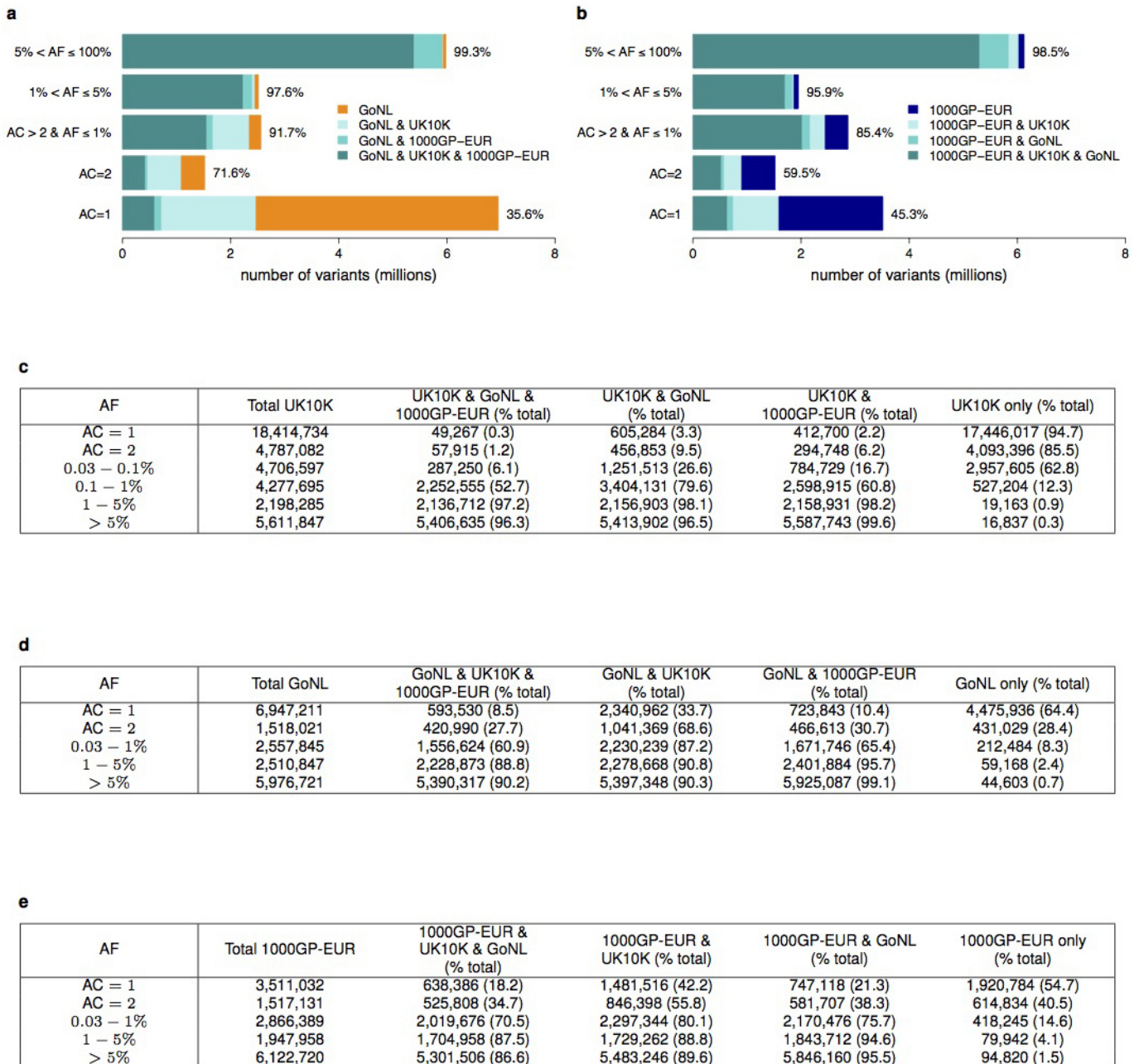
| AF | Number of variants | |
|-----------|--------------------|-----------|
| | SNVs | INDELs |
| AC = 1 | 19,596,845 | 1,132,608 |
| AC = 2 | 4,890,329 | 301,105 |
| 0.03 – 1% | 9,719,435 | 857,840 |
| 1 – 5% | 2,332,114 | 415,735 |
| > 5% | 5,787,895 | 1,479,767 |

k

| AF | WGS versus Exomes | | | | | | MZ Twins | |
|---------|-----------------------------|----------------------------|-------------|-----------------------|---------------------------|---------|-----------------------------|----------------------------|
| | Total sites (concordant, %) | Non-Ref genotypes (NRD, %) | FP (FDR, %) | FP in 1000GP (FDR, %) | FP not in 1000GP (FDR, %) | FNR (%) | Total sites (concordant, %) | Non-ref genotypes (NRD, %) |
| AC = 1 | 2,963 (99.999) | 2,965 (0.1) | 125 (4.0) | 11 (3.8) | 114 (4.1) | n.a. | 411,583 (99.995) | 3,534 (12.7) |
| AC = 2 | 1,566 (99.998) | 1,577 (0.1) | 147 (8.6) | 25 (7.9) | 122 (8.7) | n.a. | 101,116 (99.989) | 1,594 (15.1) |
| 0.03–1% | 16,303 (99.928) | 21,114 (3.3) | 1,160 (6.6) | 766 (5.5) | 394 (11.3) | 27.2 | 193,531 (99.954) | 19,034 (10.2) |
| 1 – 5% | 16,356 (99.829) | 53,165 (3.2) | 1,038 (6.0) | 980 (5.7) | 58 (68.2) | 6.4 | 50,360 (99.776) | 56,554 (4.4) |
| > 5% | 37,433 (99.688) | 1,151,178 (0.6) | 2,668 (6.7) | 2,653 (6.6) | 15 (46.9) | 7.3 | 123,690 (99.574) | 1,382,934 (0.8) |

Extended Data Figure 1 | UK10K-cohorts, sequence and sample quality and variation metrics. **a–e**, Sample quality metrics for UK10K-cohorts ($n = 3,781$) where $n = 1–1,927$ corresponds to ALSPAC and 1,928 to 3,781 to TwinsUK. This sample includes all individuals passing sample quality control, including related pairs and non-European individuals that were later removed from association tests. A subset of 3,621 individuals was included in association analyses. Samples sequenced at BGI are coloured in blue and samples sequenced at Sanger are coloured in grey. **a**, Number of singletons ($AC = 1$) by sample ($\times 10^3$). **b**, Number of INDELs by sample ($\times 10^5$). **c**, Read depth (sequence coverage) by sample. **d**, Ratio of heterozygous and homozygous non-reference (=homozygous alternative) SNV genotypes (mean for females = 1.54, mean for males = 1.47). **e**, Transition to transversion ratio (Ts/Tv) by sample. **f–i**, Sequence variation metrics for UK10K-cohorts. **f**, Types of substitution ($\times 10^6$). **g**, Number of SNVs ($\times 10^6$), INDELs ($\times 10^5$) and large deletions ($\times 10^3$) by non-overlapping non-reference allele frequency (AF) bins. **h**, Size distribution of INDELs. Negative INDEL lengths represent deletions and positive INDEL lengths represent insertions. **i**, Large deletion size distribution in unequal bin sizes where the smallest deletions were 200 bp to 1 kb long and the largest deletions 100 kb to 1 Mb. In total 18,739 deletions were called with GenomeSTRiP¹⁴. The average deletion size was ~ 13 kb and the

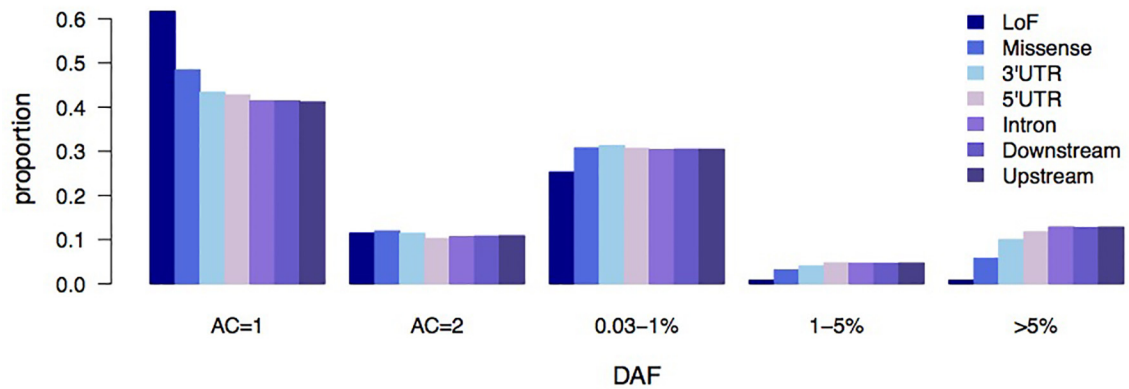
median size was ~ 3.7 kb. **j**, Total number of SNVs and INDELs by AF bin (based on 3,781 samples), multi-allelic variants are treated as separate variants. **k**, Sequence quality and variation metrics for UK10K-cohorts. For 61 overlapping TwinsUK individuals we compared the variant sites and genotypes of the low-coverage sequences with high-coverage exome data by non-overlapping AF bins (WGS versus Exomes). We considered 74,621 shared sites in non-overlapping AF bins. We calculated the fraction of concordant over total sites, the number of non-reference genotypes and non-reference genotype discordance (NRD, in %) between WGS and Exomes; false discovery rate ($FDR = FP/(FP + TP)$); TP, true positive; FP, false positive), where we consider the exomes as the truth set; number of false positives (FP) and FDR for sites that are or not shared with the 1000 Genomes Project, phase I (1000GP); false negative rate ($FNR = FN/(FN + TP)$); FN, false negative; TP, true positive), where AF bins were defined based on the 61 exomes. Furthermore, we compared 22 monozygotic twin pairs at 880,280 bi-allelic SNV sites on chromosome 20, reporting the percentage of concordant genotypes, non-reference genotypes and NRD. AFs are from the set of 3,621 samples, which contains at most one of the two monozygotic twins from each pair. We note that discrepancies can be caused by errors in either twin, so the expected NRD to the truth would be half the NRD value given.



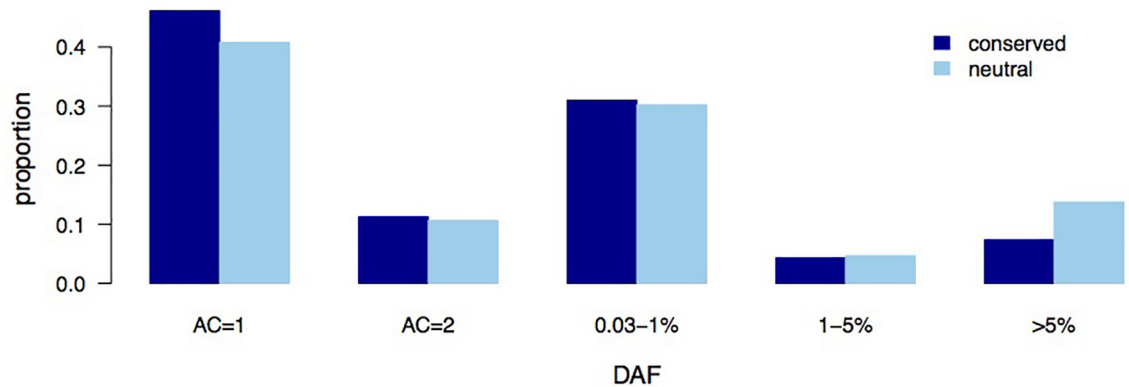
Extended Data Figure 2 | UK10K-cohorts, comparison with GoNL and 1000GP-EUR. Percentage of autosomal SNVs that are either shared between UK10K ($n = 3,781$), GoNL ($n = 499$) and 1000GP-EUR ($n = 379$), or unique to each set, for allele counts (AC) AC = 1, AC = 2, and non-overlapping allele frequency (AF) bins for higher AC. **a**, Shared and unique variants for GoNL with AF based on GoNL, and **b**, for 1000GP-EUR. AF bins are not

directly comparable owing to the different sample sizes in each call set. The x -axis shows the number of variants in millions. The percentages next to the bars represent the percentage of variants from GoNL (**a**) and 1000GP-EUR (**b**) that are shared with at least one of the other data sets. All numerical values used in **a** can be found in **d** and for **b** in **e**. **c**, Numerical values for Fig. 1.

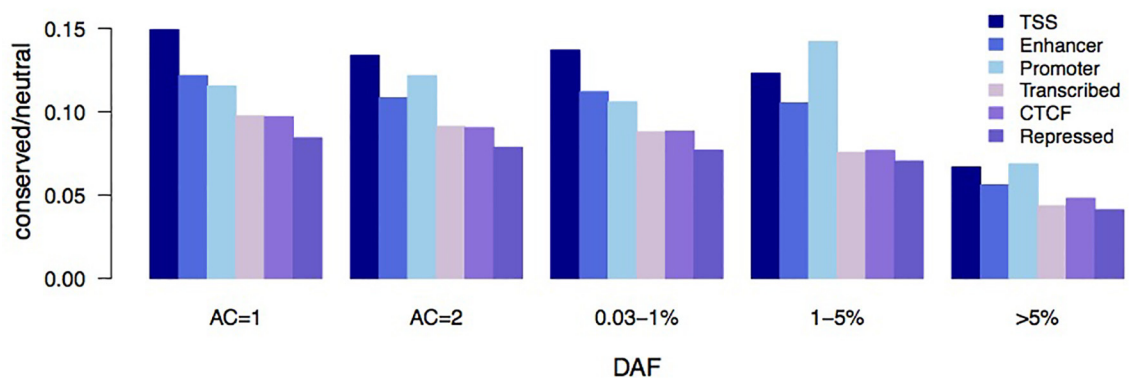
a



b

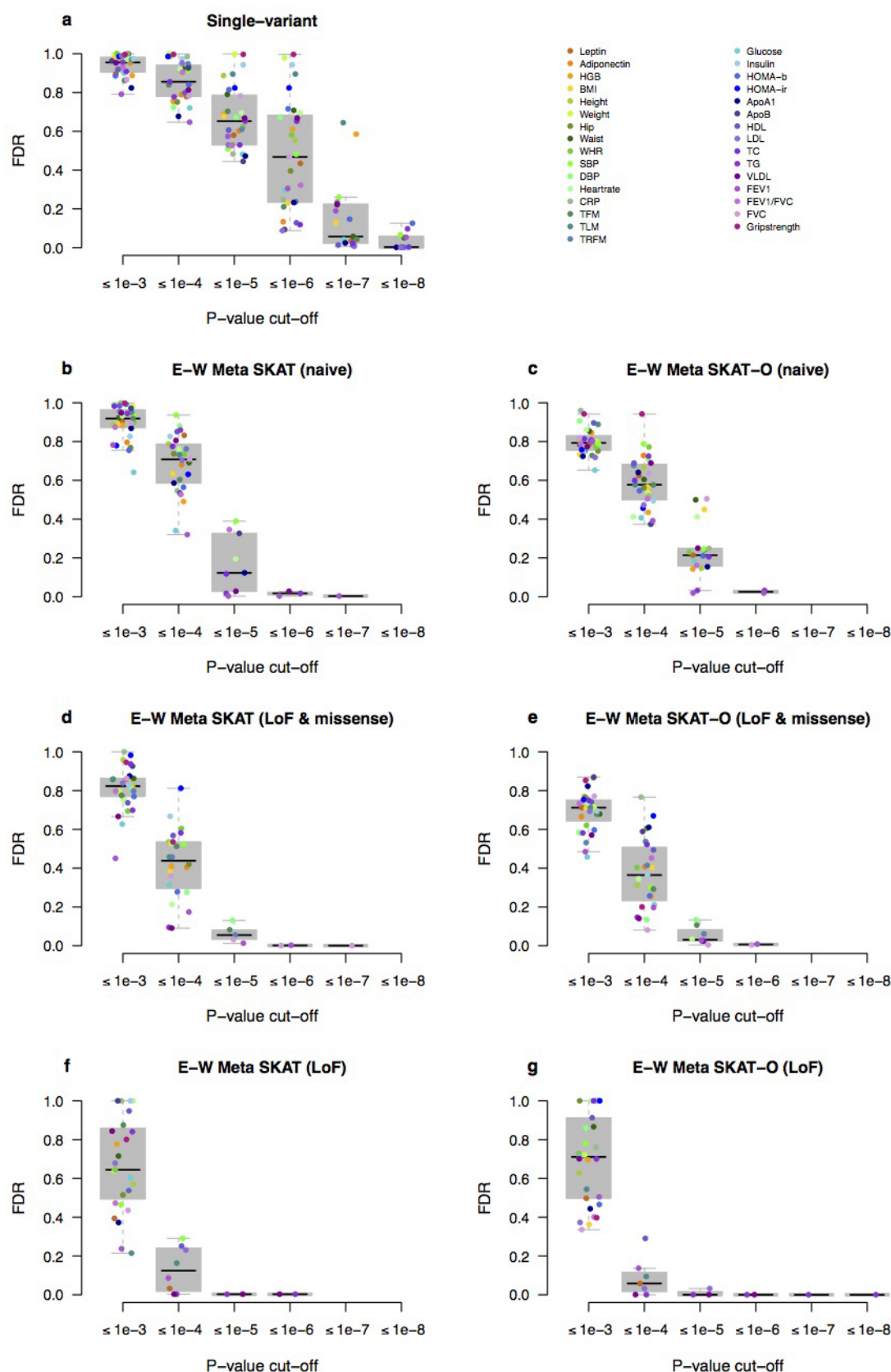


c



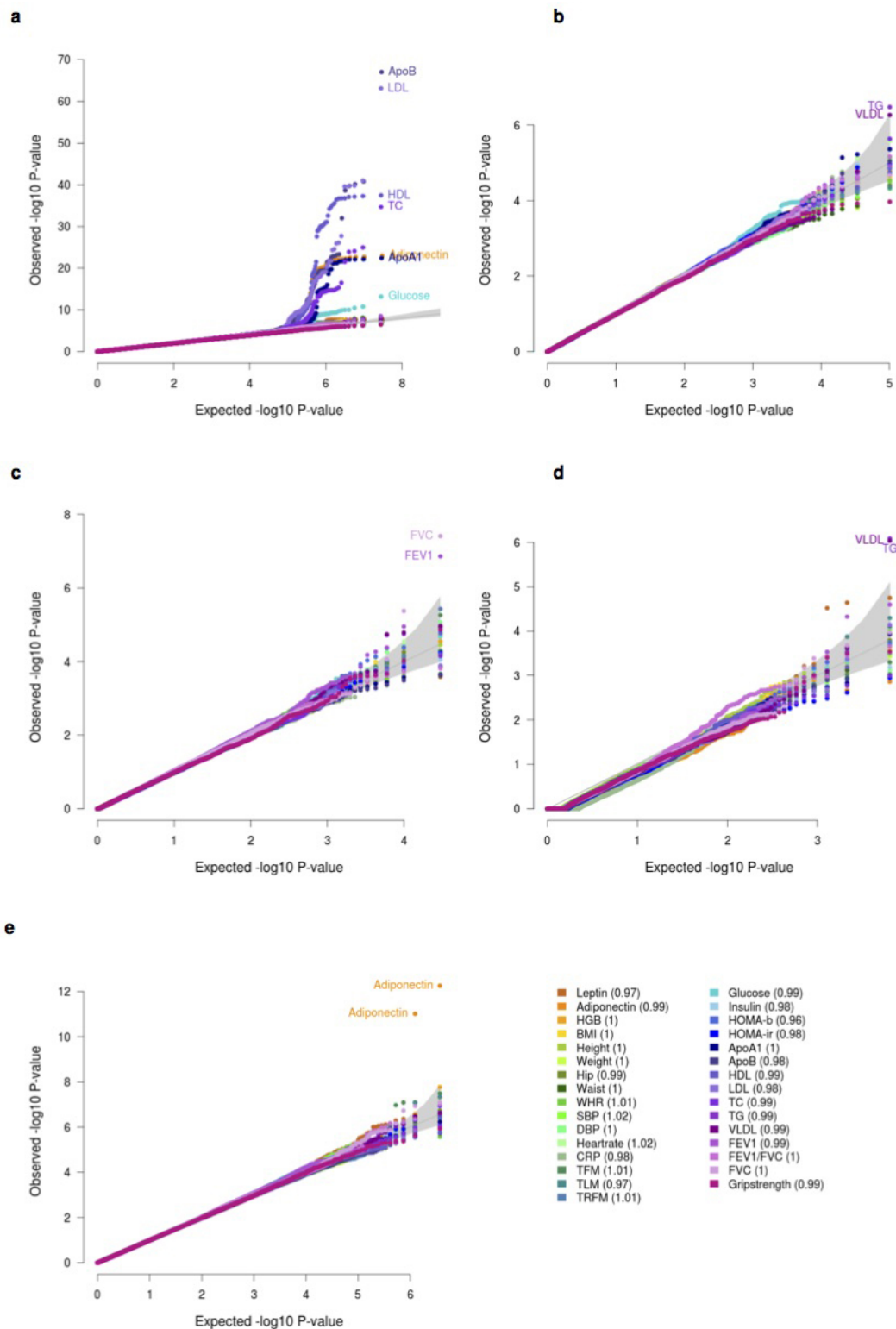
Extended Data Figure 3 | UK10K-cohorts, derived allele frequency spectrum by functional annotation. Derived allele frequency (DAF) spectrum for UK10K-cohorts chromosome 20 variants divided by functional class. **a**, Proportion of total variants (standardized across DAF bins) as a function of DAF for different genic elements. **b**, Standardized proportion of all

variants by DAF bin, and divided into conserved ($GERP > 2$) versus neutral ($GERP \leq 2$) sites. **c**, Ratio of conserved versus neutral variants by DAF bin, and classified by chromatin segmentation domains defined by ENCODE as detailed in the methods.



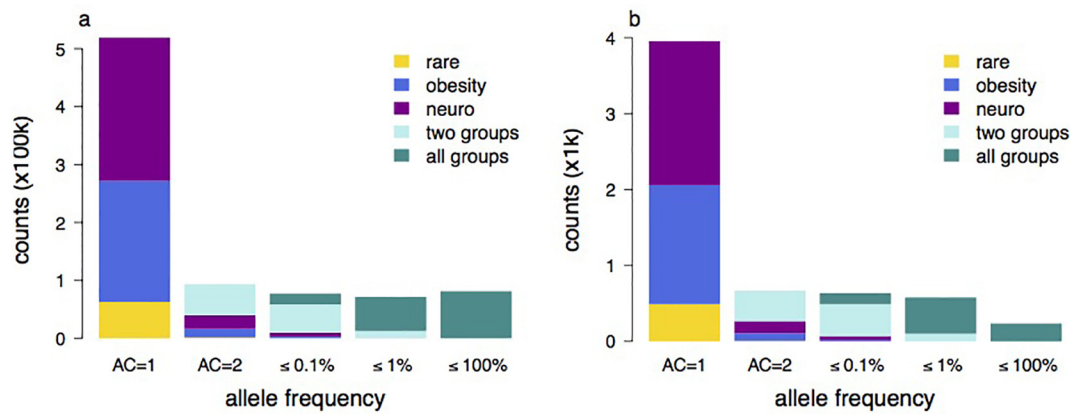
Extended Data Figure 4 | UK10K-cohorts, false discovery rate (FDR).
a–g, FDR values for reporting associations at different *P* value cut-offs for all analyses reported in this study and the 31 core traits for single-variant analysis (**a**); naive exome-wide Meta SKAT (**b**); naive exome-wide Meta SKAT-O

(**c**); functional exome-wide Meta SKAT (LoF and missense) (**d**); functional exome-wide Meta SKAT-O (LoF and missense) (**e**); functional exome-wide Meta SKAT (LoF) (**f**); functional exome-wide Meta SKAT-O (LoF) (**g**).



Extended Data Figure 5 | UK10K-cohorts, QQ plots. QQ plots for the association tests of the 31 core traits in the WGS data set ($n = 3,621$ individuals). **a**, Single-variant analysis (~ 14 million variants with $\text{MAF} \geq 0.1\%$); **b**, naive exome-wide Meta SKAT (1,783,548 variants with $\text{MAF} < 1\%$ in 50,717 windows); **c**, functional exome-wide Meta SKAT

(LoF and missense; 256,733 variants with $\text{MAF} < 1\%$ in 14,909 windows); **d**, loss-of-function functional exome-wide Meta SKAT (LoF; 9,113 variants with $\text{MAF} < 1\%$ in 3,208 windows); **e**, genome-wide Meta SKAT (35,858,684 variants with $\text{MAF} < 1\%$ in 1,845,982 windows).

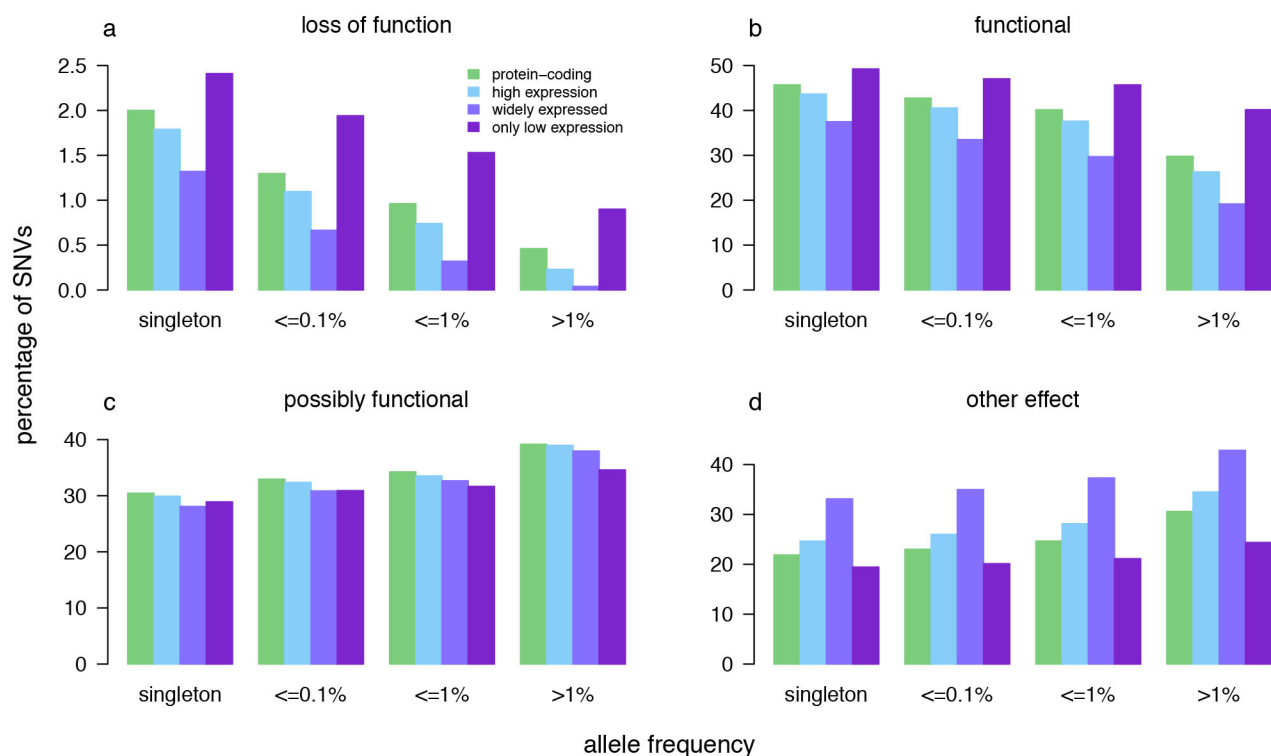


| | | Disease Collection | | | | | |
|--------|--------------------|--------------------|------------|------------|--------|---------|--------------------|
| AF | | Total | All groups | Two groups | Rare | Obesity | Neurodevelopmental |
| SNVs | AC = 1 | 518,966 | 0 | 0 | 62,989 | 208,700 | 247,277 |
| | AC = 2 | 93,601 | 0 | 54,213 | 1,392 | 15,103 | 22,893 |
| | AC = 2 < AF ≤ 0.1% | 77,199 | 18,644 | 49,140 | 86 | 3,008 | 6,321 |
| | 0.1 < AF ≤ 1% | 71,634 | 58,958 | 12,575 | 0 | 19 | 82 |
| | AF > 1% | 81,246 | 81,242 | 4 | 0 | 0 | 0 |
| AF | | Total | All groups | Two groups | Rare | Obesity | Neurodevelopmental |
| INDELs | AC = 1 | 3,954 | 0 | 0 | 491 | 1,568 | 1,895 |
| | AC = 2 | 666 | 0 | 400 | 6 | 100 | 160 |
| | AC = 2 < AF ≤ 0.1% | 635 | 144 | 422 | 0 | 16 | 53 |
| | 0.1 < AF ≤ 1% | 578 | 477 | 101 | 0 | 0 | 0 |
| | AF > 1% | 234 | 234 | 0 | 0 | 0 | 0 |

| | | UK10K exomes | | EA ESP | |
|--|--------------------|------------------------------|-------------------------------------|-----------------------------|------------------------------------|
| AF | | Total (% shared with EA ESP) | Inside baits (% shared with EA ESP) | Total (% shared with UK10K) | Inside baits (% shared with UK10K) |
| AC = 1 AC = 2 AC = 2 < AF ≤ 0.1% AF > 0.1% Total | AC = 1 | 518,966 (21) | 273,856 (23) | 681,351 (16) | 587,453 (15) |
| | AC = 2 | 93,601 (48) | 48,863 (53) | 135,426 (41) | 112,556 (39) |
| | AC = 2 < AF ≤ 0.1% | 77,199 (74) | 38,935 (81) | 132,937 (72) | 107,041 (70) |
| | AF > 0.1% | 152,880 (89) | 69,989 (99) | 201,062 (96) | 141,215 (99) |
| | Total | 842,646 (41) | 431,643 (44) | 1,150,776 (40) | 948,265 (37) |

Extended Data Figure 6 | UK10K-exomes, sequence variant statistics. Number of variants ($\times 10^3$) that are found in one or more of the three UK10K-exomes disease data sets, as a function of allele frequency (AF) of the non-reference allele. Variants are split into allele counts (AC) AC = 1, AC = 2 and non-overlapping AF bins for AC > 2. Allele frequency is the frequency of the alternative allele. The distributions of SNVs and INDELs across frequencies and disease collections are similar, except that there is a lower proportion of INDELs with AF > 1% compared to SNVs. **a**, SNVs. Multiallelic sites are included (1.6%), and non-reference alleles at the same site are treated as separate variants. **b**, INDELs. Counts are given in **c**. **c**, Variants are classed by whether they were found in more than one disease collection or unique to a

specific group. **d**, Comparison of UK10K patient set with European-Americans individuals from the NHLBI Exome Sequencing project (EA ESP). The left panel shows the variants identified in UK10K and the percentage shared with EA ESP. Both the total number of variants and the number within the EA ESP bait regions (intersection of bait sets) are given. The right panel shows the variants identified in EA ESP and the percentage shared with UK10K. Both the total number of variants, and the number within the UK10K baits after removing any that failed UK10K quality control, are given. There is some overlap in the ranges of AC and AF for EA ESP variants because different numbers of individuals were included.



e

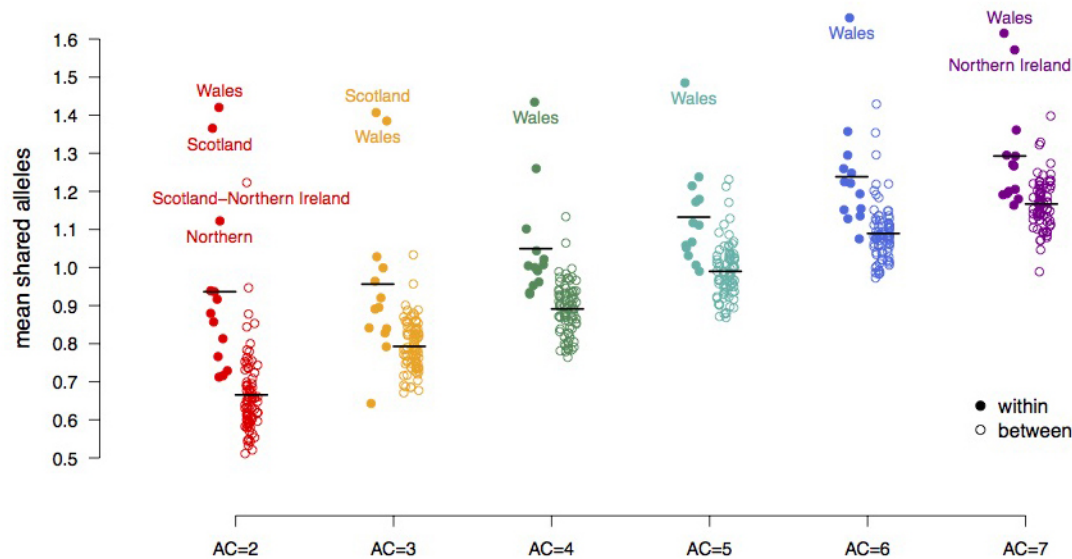
| Transcript set | Frequency | LoF | Functional | Possibly functional | Other | No qualifying transcripts |
|---------------------|--------------------|--------|------------|---------------------|---------|---------------------------|
| All | AC = 1 | 10,647 | 234,169 | 178,067 | 95,667 | 416 |
| | AC = 2 | 1,338 | 40,013 | 34,430 | 17,747 | 73 |
| | AF < 0.1% | 988 | 32,020 | 28,905 | 15,226 | 60 |
| | AF ≤ 1% | 751 | 28,264 | 27,651 | 14,900 | 68 |
| | AF > 1% | 446 | 23,721 | 36,084 | 20,860 | 135 |
| Protein-coding | AC = 1 | 10,127 | 231,825 | 154,433 | 110,928 | 11,653 |
| | AC = 2 | 1,254 | 39,538 | 29,836 | 20,634 | 2,339 |
| | AC = 2 < AF ≤ 0.1% | 905 | 31,613 | 25,007 | 17,661 | 2,013 |
| | AF ≤ 1% | 664 | 27,876 | 23,776 | 17,124 | 2,194 |
| | AF > 1% | 361 | 23,123 | 30,413 | 23,747 | 3,602 |
| High expression | AC = 1 | 6,343 | 154,247 | 105,711 | 87,097 | 56,948 |
| | AC = 2 | 760 | 26,062 | 20,365 | 16,097 | 10,846 |
| | AC = 2 < AF ≤ 0.1% | 504 | 20,663 | 16,946 | 13,866 | 9,167 |
| | AF ≤ 1% | 352 | 17,981 | 16,023 | 13,450 | 9,007 |
| | AF > 1% | 123 | 13,787 | 20,422 | 18,073 | 12,831 |
| Widely expressed | AC = 1 | 1,304 | 36,957 | 27,694 | 32,654 | 311,737 |
| | AC = 2 | 131 | 5,909 | 5,301 | 5,863 | 56,926 |
| | AC = 2 < AF ≤ 0.1% | 76 | 4,525 | 4,308 | 5,030 | 47,207 |
| | AF ≤ 1% | 40 | 3,684 | 4,050 | 4,629 | 44,410 |
| | AF > 1% | 5 | 2,440 | 4,834 | 5,456 | 52,501 |
| Only low expression | AC = 1 | 1,093 | 22,305 | 13,092 | 8,805 | 365,051 |
| | AC = 2 | 171 | 4,047 | 2,608 | 1,681 | 65,623 |
| | AC = 2 < AF ≤ 0.1% | 133 | 3,320 | 2,232 | 1,469 | 53,992 |
| | AF ≤ 1% | 104 | 3,113 | 2,157 | 1,439 | 50,000 |
| | AF > 1% | 83 | 3,705 | 3,192 | 2,249 | 56,007 |

Extended Data Figure 7 | UK10K-exomes, functional consequences.

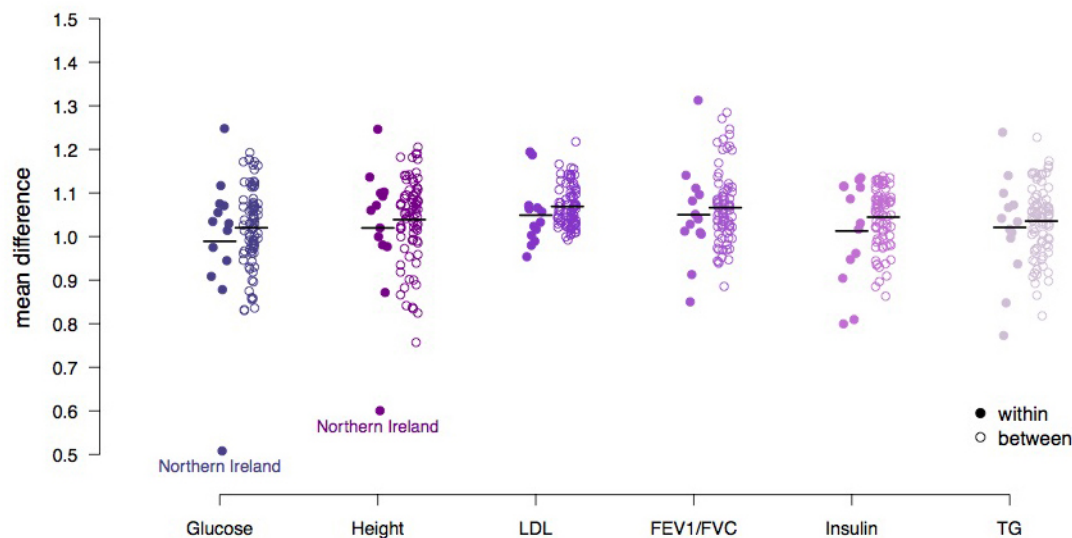
a–d, Percentage of SNVs in each allele frequency bin that are loss of function (**a**), functional (**b**), possibly functional (**c**) and other (**d**), when consequences are restricted to given subsets of transcripts, and where the most severe consequence in qualifying transcripts is used. Values are percentages of SNVs that have transcripts of a given type. Protein-coding is transcripts with a biotype of protein coding. High expression is transcripts with FPKM (fragments per kilobase of transcript per million mapped reads) ≥ 1 in any tissue. Widely expressed is transcripts with FPKM ≥ 1 in 16 tissues. Only low expression is transcripts expressed at FPKM < 1 in all 16 tissues where there were no

transcripts with high expression in that variant. Expression was determined from the Illumina Body Map data set. Variants mapping to protein-coding transcripts <300-bp long or with missing or low quality expression data were excluded. Frequency bins are singletons and non-overlapping allele frequency ranges for allele counts above 1. Allele frequency is the frequency of the alternative allele. Multi-allelic sites were included with alternative alleles at the same site treated as separate variants. **e**, Counts of single nucleotide polymorphisms in each consequence class by allele frequency and transcript subset.

a



b

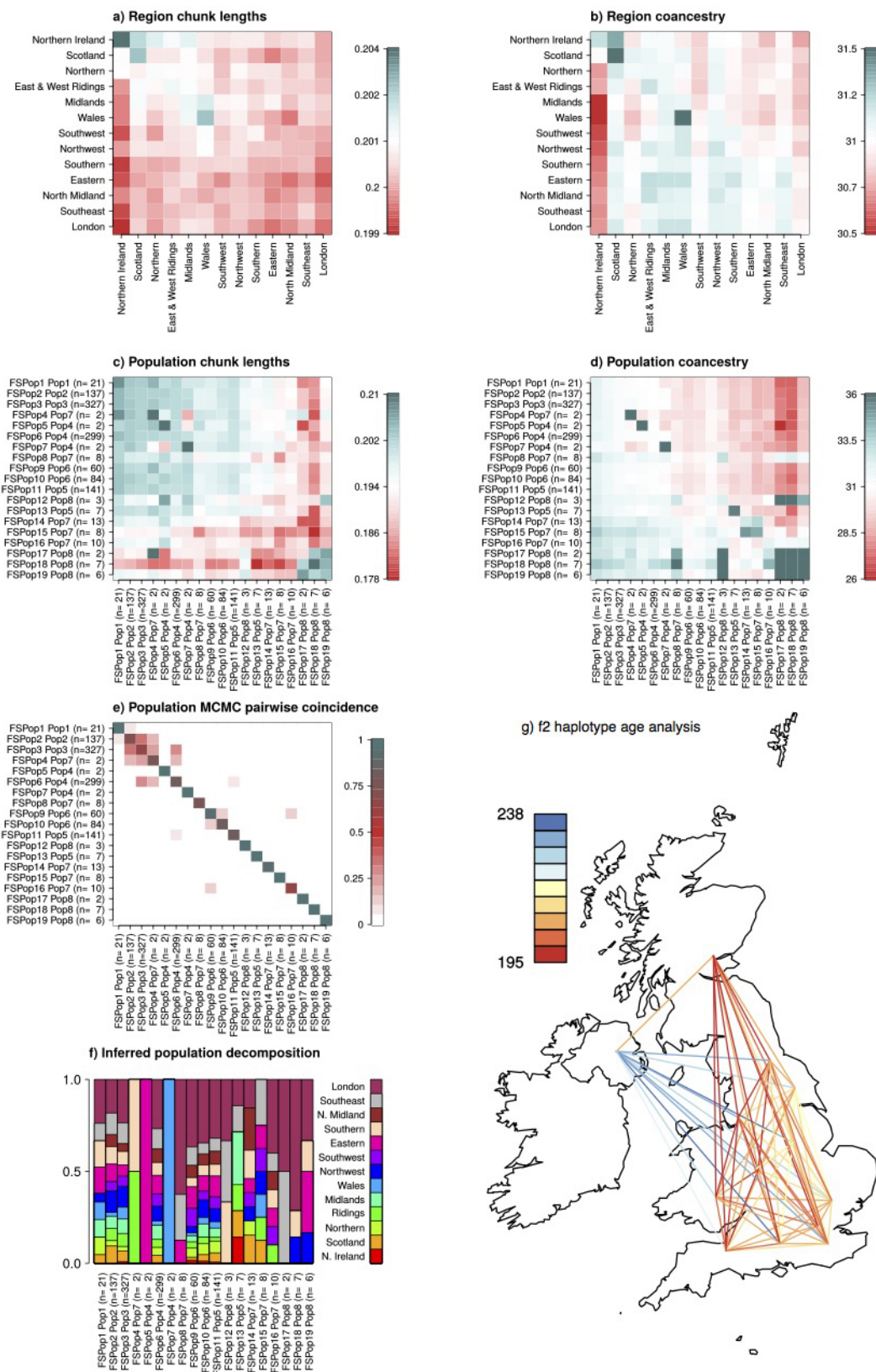


c

| AC=2 | | AC=3 | | AC=4 | | AC=5 | | AC=6 | | AC=7 | |
|-------------|-------|-------------|-------|--------------|-------|--------------|-------|-------------|-------|--------------|-------|
| Height | 0.048 | Height | 0.052 | Adiponectin | 0.190 | Gripstrength | 0.137 | FEV1/FVC | 0.142 | Insulin | 0.100 |
| LDL | 0.063 | Weight | 0.055 | TRFM | 0.231 | Adiponectin | 0.206 | Adiponectin | 0.144 | ApoA1 | 0.108 |
| Adiponectin | 0.071 | Adiponectin | 0.075 | Insulin | 0.297 | ApoB | 0.298 | Height | 0.144 | Gripstrength | 0.119 |
| Weight | 0.177 | FEV1/FVC | 0.117 | Weight | 0.317 | Insulin | 0.310 | Glucose | 0.144 | TFM | 0.175 |
| Waist | 0.192 | Waist | 0.183 | Gripstrength | 0.359 | ApoA1 | 0.318 | LDL | 0.150 | FEV1 | 0.206 |

Extended Data Figure 8 | UK10K-cohorts, genotype and phenotype similarities within and between regions. a, b, Dot plots show the genetic (a) and phenotypic distribution (b) of the relationships of 1,139 unrelated TwinsUK individuals by their regional place of birth. To determine the genetic relationships we used the mean number of shared alleles between two individuals within and between regions for allele counts (AC) 2 to 7, where AC is calculated from the whole data set of 3,781 samples. To determine phenotypic similarities we calculated the mean difference between the residualized

phenotypes. Genetically-related individuals are more closely related within a region than between regions, while the phenotypic distance measure has similar distributions within and between regions. The mean shared alleles increase with increasing allele count, and simultaneously the within and between distributions converge. c, The five lowest *P* values for AC 2 to 7 obtained from Mantel tests to determine similarities between genotypes and phenotypes by region. *P* values were not significant after correcting for multiple testing using the FDR method⁴⁹. Full trait names are given in Supplementary Table 1.



Extended Data Figure 9 | UK10K-cohorts, population fine structure in the TwinsUK sample. **a**, Chunk length matrix for all UK10K defined geographic regions, calculated as described in the methods. The bottom 5 regions are merged in Box 1 Figure. **b**, Coancestry matrix for all UK10K defined geographic regions, calculated as described in the methods. **c**, Chunk length matrix for all UK10K FineSTRUCTURE inferred populations, calculated as described in the methods. **d**, Coancestry matrix for all UK10K FineSTRUCTURE inferred populations. Details on calculation of these parameters are described in Methods. **e**, Pairwise coincidence matrix for the UK10K FineSTRUCTURE MCMC run, showing the fraction of the 1,000 retained iterations from the posterior in which each pair of individuals is in the same population, averaged for each pair of populations. The full posterior is extremely complex, which is indicative of a continuous admixture cline rather than discrete populations.

f, Sources distribution for the FineSTRUCTURE inferred populations with the full set of inferred populations and geographic labels. Geographic labels of London, Southeast, North Midland, Southern and Eastern are merged into South and East for Box 1 Figure. FSPop labels are given to populations inferred by FineSTRUCTURE, which are merged into the Pop labels as shown in the main Box 1 Figure. **g**, The f2 haplotype age analysis estimates the time to the most recent common ancestor (tMRCA) between the two haplotypes underlying a given observed variant of allele count 2 in all of the TwinsUK samples. The observed IBD segment length around each f2 variant estimates the tMRCA, using an explicit model parameterized by the recombination and the mutation rates. Shown is the map of the UK with all regions used in this analysis depicted by their location, and lines colour-coding the observed median tMRCA of f2 haplotypes.

Extended Data Table 1 | UK10K-cohorts, estimated variance explained by SNVs across the 31 UK10K traits shared by both cohorts

| Trait | N | HapMap 2 | | HapMap3 | | 1000GP | | UK10K | |
|---------------------------------------|-------|---------------|----------------------|---------------|----------------------|---------------|----------------------|---------------|----------------------|
| | | Beta (SE) | P-value | Beta (SE) | P-value | Beta (SE) | P-value | Beta (SE) | P-value |
| Anthropometry/obesity | | | | | | | | | |
| Height | 3,541 | 0.200 (0.087) | 0.009 | 0.201 (0.093) | 0.013 | 0.268 (0.102) | 0.002 | 0.262 (0.101) | 0.003 |
| BMI | 3,538 | 0.150 (0.088) | 0.044 | 0.191 (0.095) | 0.024 | 0.177 (0.104) | 0.040 | 0.150 (0.103) | 0.069 |
| Hip | 3,074 | 0.159 (0.100) | 0.052 | 0.191 (0.108) | 0.037 | 0.143 (0.116) | 0.100 | 0.116 (0.116) | 0.150 |
| Waist | 3,072 | 0.166 (0.098) | 0.038 | 0.197 (0.106) | 0.028 | 0.108 (0.115) | 0.163 | 0.083 (0.114) | 0.226 |
| WHR | 3,071 | 0.107 (0.097) | 0.127 | 0.159 (0.105) | 0.057 | 0.055 (0.114) | 0.310 | 0.053 (0.113) | 0.313 |
| Weight | 3,559 | 0.077 (0.087) | 0.190 | 0.089 (0.094) | 0.174 | 0.027 (0.103) | 0.396 | 0.008 (0.102) | 0.468 |
| Adiponectin | 2,325 | 0.312 (0.137) | 0.012 | 0.359 (0.145) | 0.007 | 0.407 (0.168) | 0.011 | 0.401 (0.167) | 0.011 |
| Leptin | 2,417 | 0.040 (0.125) | 0.374 | 0.057 (0.135) | 0.337 | 0.024 (0.148) | 0.437 | 0.008 (0.146) | 0.479 |
| TFM | 3,399 | 0.049 (0.089) | 0.289 | 0.084 (0.097) | 0.190 | 0.114 (0.104) | 0.126 | 0.088 (0.103) | 0.187 |
| TLM | 3,399 | 0.128 (0.090) | 0.074 | 0.150 (0.097) | 0.058 | 0.055 (0.108) | 0.305 | 0.046 (0.107) | 0.336 |
| TRFM | 3,197 | 0.000 (0.094) | 0.500 | 0.026 (0.101) | 0.398 | 0.055 (0.110) | 0.302 | 0.033 (0.109) | 0.378 |
| Diabetes biochemistry | | | | | | | | | |
| Glucose | 2,925 | 0.134 (0.106) | 0.101 | 0.129 (0.112) | 0.124 | 0.131 (0.125) | 0.146 | 0.108 (0.125) | 0.194 |
| Insulin | 2,896 | 0.210 (0.109) | 0.026 | 0.225 (0.114) | 0.022 | 0.170 (0.129) | 0.096 | 0.164 (0.128) | 0.103 |
| HOMA-B | 2,888 | 0.257 (0.109) | 0.008 | 0.321 (0.115) | 0.002 | 0.220 (0.129) | 0.044 | 0.230 (0.128) | 0.036 |
| HOMA-IR | 2,796 | 0.242 (0.113) | 0.016 | 0.262 (0.119) | 0.012 | 0.171 (0.134) | 0.105 | 0.172 (0.133) | 0.101 |
| Cardiovascular and blood biochemistry | | | | | | | | | |
| CRP | 2,046 | 0.056 (0.150) | 0.356 | 0.012 (0.161) | 0.472 | 0.000 (0.177) | 0.500 | 0.000 (0.175) | 0.500 |
| LDL | 3,191 | 0.117 (0.098) | 0.117 | 0.133 (0.105) | 0.106 | 0.100 (0.113) | 0.183 | 0.123 (0.113) | 0.134 |
| HDL | 3,210 | 0.217 (0.093) | 0.007 | 0.299 (0.100) | 7.0×10^{-4} | 0.318 (0.115) | 2.4×10^{-3} | 0.341 (0.113) | 7.9×10^{-4} |
| TC | 3,206 | 0.097 (0.093) | 0.139 | 0.119 (0.100) | 0.104 | 0.077 (0.110) | 0.235 | 0.116 (0.109) | 0.132 |
| TG | 3,202 | 0.032 (0.094) | 0.366 | 0.078 (0.101) | 0.215 | 0.000 (0.111) | 0.500 | 0.000 (0.111) | 0.500 |
| VLDL | 3,197 | 0.033 (0.094) | 0.361 | 0.081 (0.101) | 0.206 | 0.000 (0.111) | 0.500 | 0.000 (0.111) | 0.500 |
| ApoA1 | 2,914 | 0.174 (0.102) | 0.033 | 0.256 (0.109) | 0.005 | 0.176 (0.125) | 0.076 | 0.238 (0.122) | 0.019 |
| ApoB | 2,911 | 0.029 (0.107) | 0.395 | 0.071 (0.117) | 0.278 | 0.000 (0.124) | 0.500 | 0.000 (0.123) | 0.500 |
| HGB | 3,077 | 0.178 (0.099) | 0.029 | 0.220 (0.104) | 0.012 | 0.225 (0.121) | 0.030 | 0.261 (0.118) | 0.011 |
| Heart, lung function, dynamic | | | | | | | | | |
| DBP | 3,309 | 0.154 (0.093) | 0.045 | 0.174 (0.099) | 0.037 | 0.191 (0.109) | 0.034 | 0.186 (0.109) | 0.038 |
| SBP | 3,309 | 0.278 (0.094) | 0.001 | 0.310 (0.101) | 8.3×10^{-4} | 0.345 (0.112) | 7.0×10^{-4} | 0.362 (0.112) | 3.9×10^{-4} |
| Heart rate | 2,975 | 0.150 (0.104) | 0.071 | 0.180 (0.112) | 0.051 | 0.129 (0.126) | 0.156 | 0.134 (0.124) | 0.140 |
| FEV1 | 3,287 | 0.481 (0.094) | 4.7×10^{-8} | 0.534 (0.101) | 2.6×10^{-8} | 0.545 (0.114) | 5.9×10^{-7} | 0.562 (0.112) | 1.3×10^{-7} |
| FVC | 3,285 | 0.420 (0.094) | 1.6×10^{-6} | 0.467 (0.101) | 9.9×10^{-7} | 0.479 (0.114) | 9.8×10^{-6} | 0.487 (0.113) | 5.0×10^{-6} |
| FEV1/FVC | 3,280 | 0.294 (0.093) | 5.1×10^{-4} | 0.335 (0.101) | 3.4×10^{-4} | 0.361 (0.111) | 3.2×10^{-4} | 0.367 (0.110) | 2.1×10^{-4} |
| Grip strength | 3,196 | 0.270 (0.096) | 0.002 | 0.248 (0.103) | 0.007 | 0.323 (0.115) | 0.002 | 0.334 (0.114) | 0.001 |

We used the restricted maximum likelihood (REML) method implemented in GCTA to estimate phenotypic variance explained by SNV sets ($MAF \leq 1\%$) in our discovery sequence data ($n = 3,621$ individuals). SNVs were selected from the WGS data to correspond to the content of four different reference panels: HapMap2 ($n = 2,331,713$ SNVs), Hapmap3 ($n = 1,168,695$), 1000 Genomes ($n = 7,475,230$) and the entire UK10K reference panel ($n = 8,317,582$). Each GRM was individually tested against the 31 traits with phenotypic values present in both cohort studies, producing a beta, s.e. and P value for total trait variance explained by the given SNV set. Full trait names are given in Supplementary Table 1.

A subthermionic tunnel field-effect transistor with an atomically thin channel

Deblina Sarkar¹, Xuejun Xie¹, Wei Liu¹, Wei Cao¹, Jiahao Kang¹, Yongji Gong², Stephan Kraemer³, Pulickel M. Ajayan² & Kaustav Banerjee¹

The fast growth of information technology has been sustained by continuous scaling down of the silicon-based metal-oxide field-effect transistor. However, such technology faces two major challenges to further scaling. First, the device electrostatics (the ability of the transistor's gate electrode to control its channel potential) are degraded when the channel length is decreased, using conventional bulk materials such as silicon as the channel. Recently, two-dimensional semiconducting materials^{1–7} have emerged as promising candidates to replace silicon, as they can maintain excellent device electrostatics even at much reduced channel lengths. The second, more severe, challenge is that the supply voltage can no longer be scaled down by the same factor as the transistor dimensions because of the fundamental thermionic limitation of the steepness of turn-on characteristics, or subthreshold swing^{8,9}. To enable scaling to continue without a power penalty, a different transistor mechanism is required to obtain subthermionic subthreshold swing, such as band-to-band tunnelling^{10–16}. Here we demonstrate band-to-band tunnel field-effect transistors (tunnel-FETs), based on a two-dimensional semiconductor, that exhibit steep turn-on; subthreshold swing is a minimum of 3.9 millivolts per decade and an average of 31.1 millivolts per decade for four decades of drain current at room temperature. By using highly doped germanium as the source and atomically thin molybdenum disulfide as the channel, a vertical heterostructure is built with excellent electrostatics, a strain-free heterointerface, a low tunnelling barrier, and a large tunnelling area. Our atomically thin and layered semiconducting-channel tunnel-FET (ATLAS-TFET) is the only planar architecture tunnel-FET to achieve subthermionic subthreshold swing over four decades of drain current, as recommended in ref. 17, and is also the only tunnel-FET (in any architecture) to achieve this at a low power-supply voltage of 0.1 volts. Our device is at present the thinnest-channel subthermionic transistor, and has the potential to open up new avenues for ultra-dense and low-power integrated circuits, as well as for ultra-sensitive biosensors and gas sensors^{18–21}.

Two-dimensional (2D) semiconducting materials derived from transition metal dichalcogenides (TMDs) are highly promising as channel material for FETs, specifically for mitigating the degradation of device electrostatics (Supplementary Information S1), and hence have attracted much attention recently^{1–7}. Their ultra-thin structure and pristine interfaces can lead to excellent electrostatics, and at the same time their planar nature facilitates easy fabrication compared to one-dimensional structures (such as nanowires and nanotubes) that can also provide excellent electrostatics but are far less fabrication-friendly.

Even if excellent electrostatics could be achieved with atomically thin TMDs, the most severe challenge for metal-oxide-semiconductor field-effect transistors (MOSFETs) still remains, which is the increase in power density due to the inability to scale down the supply voltage. This arises from the fundamental thermionic limitation of the subthreshold swing (SS) of $2.3k_B T/q$ (or 60 mV per decade at room temperature) in conventional FETs (CFETs, Supplementary Information

S2); here k_B is Boltzmann's constant, T is temperature and q is the elementary charge. (SS is the inverse of the subthreshold slope and is given by $SS = (d\log_{10} I_{DS}/dV_{GS})^{-1}$, where I_{DS} is the drain-to-source current, and V_{GS} the gate-to-source voltage.) Thus, just using 2D semiconducting-channel materials only partially addresses the scaling issue—use of novel device technology based on 2D materials is necessary for simultaneous achievement of efficient electrostatics as well as novel carrier transport mechanisms, in order to achieve SS values below 60 mV per decade and thereby combat power density increase and enable scaling to continue in future. Apart from digital electronics, achievement of a device based on 2D semiconducting-channel material and with subthermionic SS would be highly desirable for next-generation ultra-sensitive, low-power and fast biosensors and gas sensors. Here we demonstrate planar transistors based on a 2D semiconducting material (bilayer molybdenum disulfide, MoS₂), which overcomes the fundamental limitation on SS of CFETs, and offers a minimum SS of 3.9 mV per decade and an average SS of 31.1 mV per decade for more than four decades of drain current at room temperature. This is achieved by using a fundamentally different transport mechanism in the form of quantum mechanical band-to-band tunnelling (BTBT)^{22,23}.

Tunnel-FETs (TFETs) using BTBT^{10–16} are promising candidates for the achievement of subthermionic SS. In spite of the high level of interest in TFETs with 2D channel materials, and experimental work^{24,25} in this direction using electrostatic doping techniques, there has not until now been a successful experimental demonstration of a TFET—or of any transistor based on a 2D material with subthermionic SS. Moreover, use of electrostatic doping requires an extra gate electrode for functioning and hence, is not energy-efficient. Here we build a unique vertical TFET with subthermionic SS by engineering the substrate, portions of which are configured as a highly doped semiconductor source and other portions of which are etched and filled with a dielectric for hosting the drain and gate metal contacts, while ultra-thin 2D TMD forms the channel (Fig. 1a). This TFET structure offers several unique advantages. First, the use of a 2D TMD material as the channel produces not only excellent electrostatics but also a small tunnelling distance or tunnelling barrier width (which is determined by the channel thickness), as needed to increase the BTBT current. We note that using a 3D material as the source does not hamper device electrostatics, as it is the channel region that needs to be modulated by the gate and it is atomically thin in our case. Second, combining 3D and 2D materials opens up unprecedented opportunities for designing custom-built heterostructures. We have chosen germanium (Ge) as the 3D material because it has a relatively low electron affinity (EA) and bandgap compared to the other commonly used group IV and III-V semiconductors, while MoS₂ is chosen as the 2D material since it has a relatively high EA compared to the other commonly explored TMDs. Thus, Ge-MoS₂ forms a staggered heterojunction (Fig. 1b), with a small band overlap at the interface, leading to low tunnelling barrier height, which is necessary for increasing the BTBT

¹Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106, USA. ²Department of Materials Science and Nanoengineering, Rice University, Houston, Texas 77005, USA. ³Department of Materials, University of California, Santa Barbara, California 93106, USA.

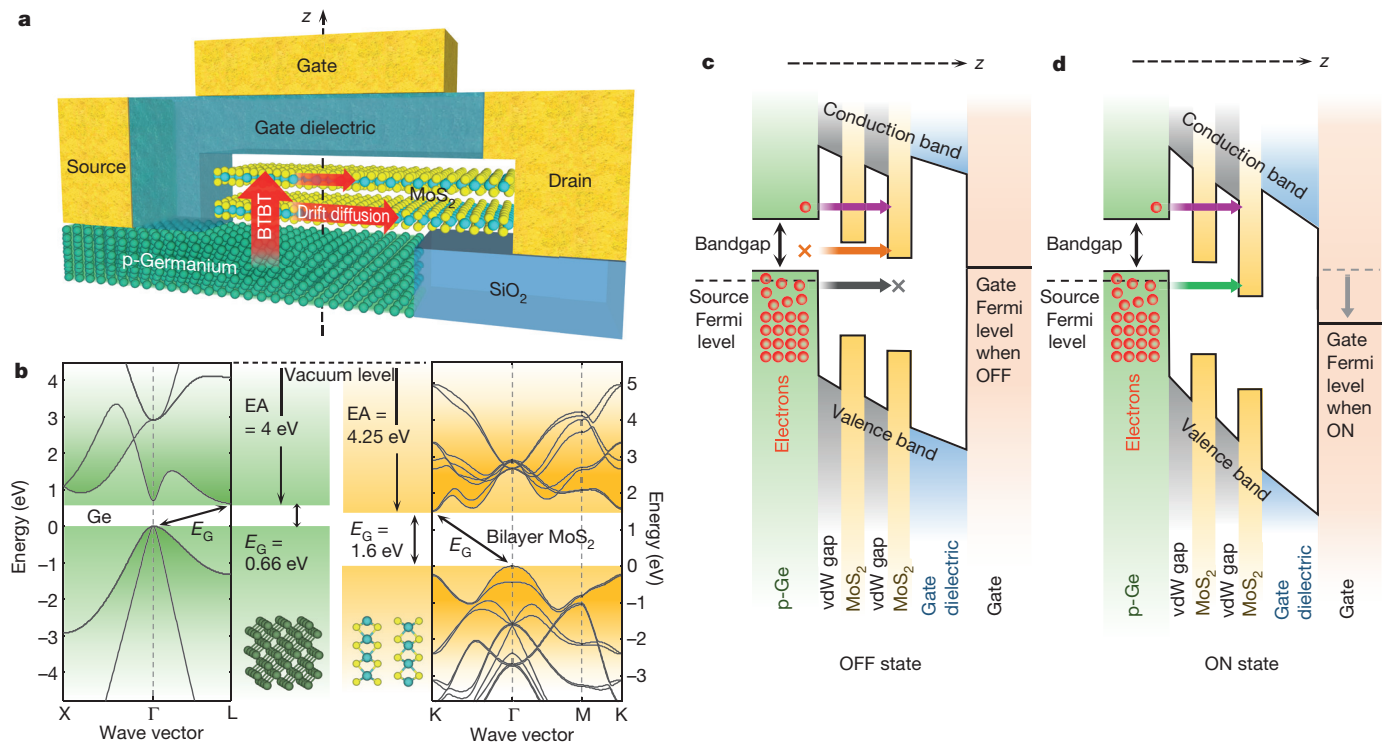


Figure 1 | Schematic and working principle of the ATLAS-TFET.

a, Schematic diagram illustrating the cross-sectional view of the ATLAS-TFET with ultra-thin bilayer MoS₂ (1.3 nm) as the channel and degenerately doped p-type Ge as the source. Path for electron transport is shown by the red arrows, which run vertically (indicating band-to-band-tunnelling, BTBT) from the Ge source to the MoS₂ and then laterally through the MoS₂ layers (via drift diffusion) to the drain. As the Ge is highly doped, the tunnelling barrier height is mainly determined by the effective band overlap between Ge and MoS₂ while the tunnelling width is determined by the MoS₂ thickness (including the van der Waals gap). **b**, Band alignment of Ge and bilayer MoS₂ showing their electron affinities (EA) and bandgaps (E_G) and, thus, illustrating the formation of a staggered vertical heterojunction. Insets in the middle show the crystal structures of both materials, while the bandstructures are shown on the left and right sides. **c**, **d**, Band diagrams along the vertical dashed line in **a** are shown in both OFF (**c**) and ON (**d**) states. The white regions represent the forbidden gaps (zero density of states, DOS). While the effective bandgap of bilayer MoS₂ is illustrated in **b**, here the bands for the two layers are shown

separately with the van der Waals (vdW) gap between them for better visual interpretation of current flow. Note that the drain contact is located perpendicular to the plane of the figure and is not shown. In the OFF state, electrons from the valence band of Ge cannot transport to MoS₂ owing to the non-availability of DOS in MoS₂ (horizontal black arrow and cross sign). At higher energies, empty DOS is available in MoS₂, but no DOS is available in Ge, again forbidding electron flow (horizontal orange arrow and cross sign). With a further increase in energy reaching above the conduction band of Ge, DOS is available in both Ge and MoS₂. However, the number of electrons available in the conduction band of the Ge source is negligible owing to the exponential decrease in electron concentration with increase in energy above the Fermi level according to the Boltzmann distribution. Thus, very few electrons can flow to the MoS₂ (horizontal purple arrow), leading to a very low OFF-state current. With an increase in gate voltage (**d**), the conduction band of MoS₂ at the dielectric interface is lowered below the valence band of the Ge source, and electrons start to flow (horizontal green arrow), resulting in an abrupt (subthermionic) increase in BTBT current.

current. Third, the heterojunction is formed with van der Waals bonds and thus has strain-free interfaces. Fourth, although methodologies for obtaining stable (and high) doping in TMDs are very challenging and still under investigation, 3D materials already enjoy well developed doping technologies that have been used in this work to form a highly doped source. This enables the creation of an ultra-sharp doping profile and hence, a high electric field at the source-channel interface, because there is a negligible chance of diffusion of dopant atoms across the heterojunction owing to the presence of a van der Waals gap. Last, because MoS₂ is placed on top of Ge forming a vertical source-channel junction, BTBT can take place across the entire area of MoS₂-Ge overlap, which leads to a higher current in the ON state than in the case of line overlap obtained in lateral junctions.

Although we are using the term tunnel-FET in a general way, our device is specifically a band-to-band tunnel-FET, involving transition of carriers from the valence band (of Ge) to the conduction band (of MoS₂). Although ‘tunnelling-transistors’ using heterostructures of 2D materials have been reported^{26,27}, they did not involve BTBT and hence, cannot lead to devices with SS below 60 mV per decade, because of the fundamental inability of a single carrier tunnelling barrier to provide this (Supplementary Information S3).

Our ATLAS-TFET provides several beneficial attributes relative to other subthermionic transistors. The use of bilayer MoS₂, which is only

1.3 nm thick, leads to a very thin channel transistor with subthermionic SS, which can lead to opportunities for ultra-dense and low-power electronic applications. We have achieved this on a planar platform, which is easily manufacturable compared to 1D structures such as nanowires and nanotubes. It is noteworthy that the International Technology Roadmap for Semiconductors (ITRS) has prescribed the attainment of average SS lower than 60 mV per decade over four decades of current. The only experimental TFET so far reported in the literature to have obtained this metric was produced by Tomioka *et al.*²⁸, who used a 1D (nanowire) based structure. The ATLAS-TFET is the first TFET demonstrated in planar architecture to satisfy this ITRS prescription, and is the only one to achieve it in any architecture at an ultra-low drain-source voltage V_{DS} of 0.1 V, which is highly desirable for the lowering of supply voltage and hence, power dissipation.

Figure 1c and d demonstrates the operation of the ATLAS-TFET using band diagrams obtained along the vertical dashed line in Fig. 1a, in both the OFF (Fig. 1c) and the ON (Fig. 1d) state. Our device is an n-type transistor, in which positive voltage is applied to the drain electrode (with respect to the source electrode) contacting the MoS₂ layers, which in turn contact the highly p-doped Ge source. Hence, electrons tend to move from the Ge to the MoS₂ and this electron transport can be modulated by the gate to turn the device ON or OFF. In the OFF state, only electrons above the conduction band of

Ge can transport to MoS₂ (purple arrow), leading to ultra-low current due to the scarcity of available electrons at high energies above the Fermi level. At lower energies, no electrons can flow owing to the non-availability of density of states (DOS) in the Ge source (orange arrow) or in the MoS₂ channel (black arrow). Hence, the OFF current is very low. As the gate voltage is increased, the conduction band of MoS₂ is lowered below the valence band of the Ge source (ON state), and hence filled DOS in the source gets aligned with empty DOS in the channel, leading to an abrupt increase in electron flow (green arrow) and hence, current, which can lead to subthermionic SS. Electrons, after tunnelling from the Ge source to the MoS₂, are 'sucked in' laterally by the drain contact, as shown by the red arrows in Fig. 1a. Note that bilayer MoS₂ is used instead of a monolayer. Although bilayer MoS₂ is 0.65 nm thicker than monolayer MoS₂, the former still offers excellent electrostatics and an ultra-low tunnelling barrier width; at the same time, it has a smaller bandgap⁴, higher DOS and is more robust to surface scattering²⁹.

For fabricating the ATLAS-TFET, we first prepare the engineered substrate. We start with a degenerately p-doped Ge wafer and etch ~300-nm-deep trenches in it, followed by the filling up of the trenches with SiO₂ dielectric and subsequent planarization. Next, formation of 20

nm/50 nm Ni/Au source contact to Ge, as well as cross marks and numberings (to assist electron-beam lithography in subsequent steps), are carried out in a single step. The engineered substrate with the source contact pads and markings is shown in Fig. 2a. To enable scalable technology, MoS₂ is synthesized using the chemical vapour deposition (CVD) process (details in Supplementary Information S4) and transferred onto the engineered substrate (Fig. 2b), followed by etching of MoS₂ except in regions where we plan to make the devices (Fig. 2c). Subsequently, the drain contact to MoS₂ is defined using electron-beam lithography followed by metallization with 20 nm/50 nm Y/Au (Fig. 2d). The TMD technology is still undergoing development, and formation of scaled and high-quality gate dielectrics remains a challenge. Hence we form the gate capacitor of the TFET, not from a conventional high-k dielectric, but from a solid polymer electrolyte, consisting of poly(ethylene oxide) and lithium perchlorate (LiClO₄), which can lead to high gate capacitance due to the formation of an electrical double layer³⁰. With advances in TMD technology and the achievement of high-quality scalable gate dielectric materials, a conventional gating method using high-k gate dielectrics could be integrated into the ATLAS-TFET. Details of the processing steps are given in Supplementary Information S5.

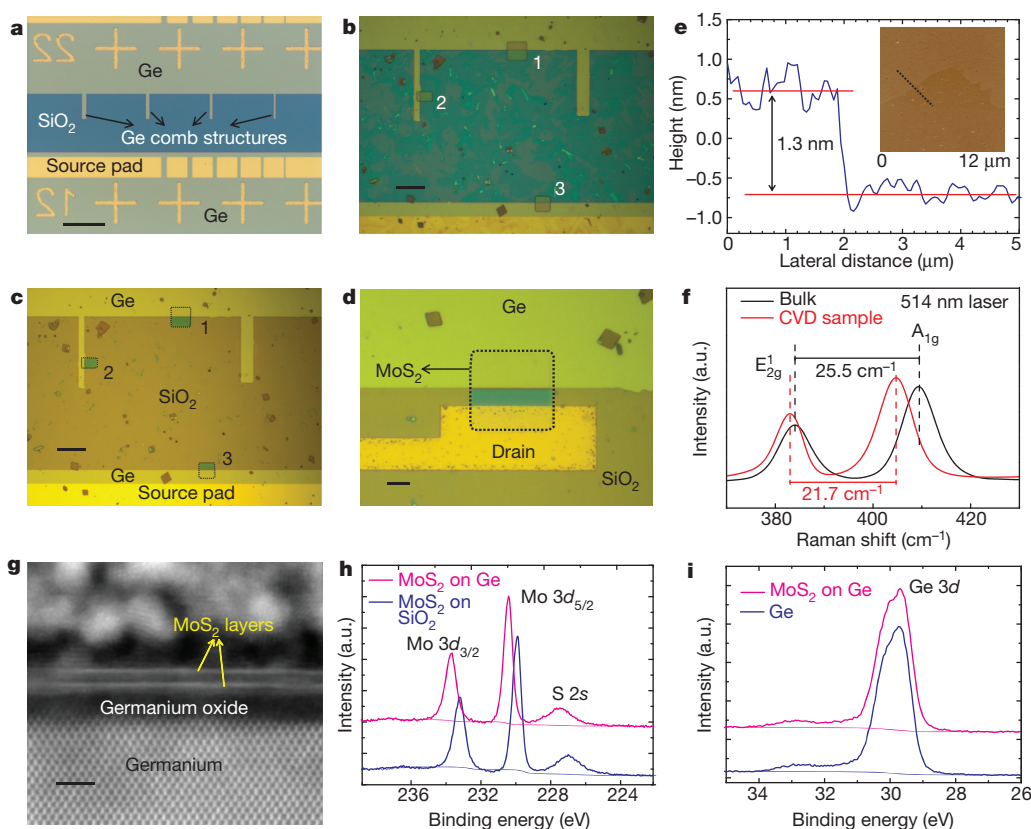


Figure 2 | Fabrication of an ATLAS-TFET and characterization results.

a, Engineered substrate consisting of alternate layers of Ge and SiO₂ along with Ge comb structures. The Ge comb structures increase the probability of achieving 'necessary overlap regions', which means that part of the MoS₂ flake overlaps the Ge while the other part overlaps the SiO₂ and thus combs are important, especially if the flakes are small. Metal source pads as well as markings required to assist electron-beam lithography are also shown. Scale bar, 100 μ m. **b**, Engineered substrate with the CVD synthesized MoS₂ transferred onto it. Regions marked 1, 2 and 3 in the image show three different ways in which 'necessary overlap regions' can be obtained. Scale bar, 25 μ m. **c**, After etching the MoS₂ from all other regions except those marked 1, 2 and 3. Scale bar, 25 μ m. **d**, After drain formation using electron-beam lithography for the region marked 1 in **b** and **c**. Scale bar, 5 μ m. More details on device fabrication can be found in Supplementary Information S5. **e**, Atomic force microscopy performed on bilayer MoS₂ confirming that the thickness is 1.3 nm.

The height profile is plotted along the dotted line shown in the inset. **f**, Raman spectroscopy (using a 514 nm laser) of the CVD bilayer sample (red curve), showing the E_{2g} peak at 383.0 cm⁻¹ and the A_{1g} peak at 404.7 cm⁻¹, which confirms that the sample is indeed bilayer MoS₂. Comparison is also shown with spectrum of bulk MoS₂ (black curve) for reference. **g**, Cross-sectional TEM image of bilayer MoS₂ on Ge. Scale bar, 2 nm. **h**, Comparison of Mo 3d core level doublet (3d_{3/2} and 3d_{5/2}) of MoS₂ on SiO₂ with that of MoS₂ on Ge, showing a 480 meV shift towards higher binding energy in the latter case. This is due to band bending in MoS₂ because of the presence of positive charges in it when placed on Ge. The sulfur 2s core level also shows a similar shift in binding energy. **i**, Comparison of the Ge 3d core level for Ge alone and for Ge with MoS₂ on it, showing negligible difference between the two, indicating an absence of band bending in Ge due to its high doping level. The deconvolution of the Ge 3d level is not shown here for clarity, and is illustrated in Supplementary Information S6.

The MoS₂ sample is characterized using atomic force microscopy and Raman spectroscopy, as shown in Fig. 2e and f, respectively. A transmission electron microscope (TEM) image of the cross-section of bilayer MoS₂ on Ge is presented in Fig. 2g, which clearly reveals the two MoS₂ layers, the Ge crystal and the presence of a thin layer of native germanium oxide. X-ray photoelectron spectroscopy (XPS) was performed to investigate the electronic structure of the Ge-MoS₂ heterostructure. As is evident from Fig. 2h, the Mo 3*d* core level in the heterostructure shifts towards higher binding energies by about 480 meV compared to that of pristine MoS₂. The sulfur 2*s* levels also shift by the same amount. This shift is due to band bending in MoS₂ in the heterostructure owing to the presence of positive charges. No shift in the Ge core levels is observed, which is consistent with the fact that band bending in Ge is almost negligible owing to its high doping (Fig. 2i). Details of the XPS results are presented in Supplementary Information S6.

The electrical characterization of the device is first performed in a two-terminal configuration, just using the source and drain contacts, without any polymer gate (Fig. 3a). We find that the device essentially behaves like a p-n junction (Fig. 3b). Note that no rectification is observed and a large current is obtained even under the reverse bias condition because of the high BTBT current due to the ultra-thin tunnelling barrier. The trend towards negative differential resistance, as shown by the circled region in the forward bias characteristics in Fig. 3b, confirms the existence of BTBT (Supplementary Information S7). BTBT is further confirmed through temperature-dependent measurements, as shown in Supplementary Information S8. Next, the transistor analysed is measured in a three-terminal configuration using the source, drain and gate (Fig. 4a). Figure 4b shows the transfer (*I*_{DS}–*V*_{GS}) characteristics of the device for different *V*_{DS} starting from a drain voltage as low as 0.1–1 V. It is observed that for all the drain voltages, the ATLAS-TFET can overcome the fundamental limitations on SS (60 mV per decade at room temperature) in MOSFETs, and SS values below 60 mV per decade are obtained over about four decades of current. We note that although the low SS occurs at a negative gate voltage, it can be adjusted by changing the work function of the gate metal. Also, the achievement of sub-60 mV per decade at room temperature using our ATLAS-TFET is repeatable, and the hysteresis in the transfer characteristics is negligible, as shown in Supplementary Information S10 and S11, respectively. The output characteristics are shown in Supplementary Information S12.

To compare the performance of our ATLAS-TFET with a CFET, a CFET is fabricated using the same MoS₂ thickness and measured under similar conditions (Supplementary Information S13). The SS of the ATLAS-TFET and the CFET is plotted in Fig. 4c. The lowest SS achieved for the CFET is 60 mV per decade, whereas for the ATLAS-TFET, not only is a minimum SS as low as 3.9 mV per decade obtained, but excellent average SS values (in mV per decade) of 5.5, 12.8, 22 and 31.1 are obtained over 1, 2, 3 and 4 decades of current, respectively. (These average values of SS have been derived using equation (s3) in Supplementary Information S3, and the data points within the range of drain currents from around 10^{–13} A to 10^{–12}/(10^{–11}/10^{–10}/10^{–9}) A are used for obtaining the average over 1/(2/3/4) decades of current.) In Fig. 4c, the noisy data points have been eliminated and thus the average of the point SS values of ATLAS-TFET plotted in that figure leads to a similar average SS over four decades, as obtained above. The performance of the ATLAS-TFET is also compared to that of other experimental TFETs with subthermionic SS reported in the literature (Supplementary Information S14). In addition to the superior SS of the ATLAS-TFET (compared to that of all other TFETs with subthermionic SS), the current obtained at a low *V*_{DS} of 0.1 V is more than two orders of magnitude larger than that obtained in ref. 28 (which is the only other TFET to have obtained subthermionic SS over four decades of drain current, although at higher *V*_{DS}) for the same *V*_{DS}. The current in the ATLAS-TFET could be further improved by removing the interfacial germanium oxide layer, which adds extra tunnelling resistance to the device as explained in Supplementary Information S15. Intrinsically, the ATLAS-TFET is promising for obtaining a high ON current because of the larger area for tunnelling (compared to the case of line overlap obtained in lateral junctions) as well as the small tunnelling barrier width, which is determined by the channel thickness (Supplementary Information S16). Higher ON current could improve the gate-to-source capacitance, while the gate-to-drain capacitance could be reduced by introducing an underlap region (Supplementary Information S17). The ATLAS-TFET is also advantageous compared to 1D TFETs, as discussed in Supplementary Information S18. Finally, to ensure that the performance of the ATLAS-TFET was not due to any substrate effects from Ge, we fabricated and characterized a CFET having identical channel (bilayer MoS₂) and gate dielectric (polymer complex) materials, and where p-Ge acts mainly as the substrate and not as the source because the source metal is directly connected to the

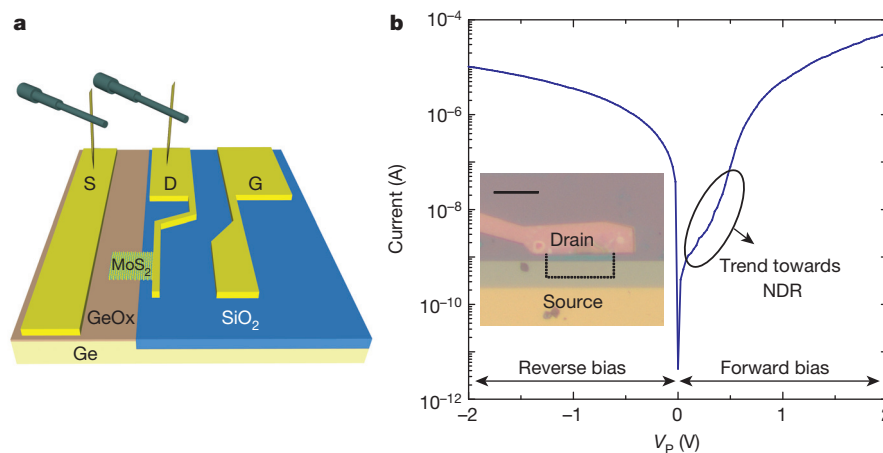


Figure 3 | Room-temperature electrical characteristics of an ATLAS-TFET in diode configuration. **a**, Schematic diagram showing the probing configuration for measurement of the characteristics of the p–n junction diode formed by highly p-doped Ge and naturally n-doped MoS₂. The presence of a thin layer of germanium oxide (GeOx) is also shown. Note that, although there is a layer of germanium oxide between the source contact and the Ge, we have still achieved ohmic contact to Ge, owing to the ultra-thin nature of the oxide. **b**, Current as a function of applied bias (*V*_p) on Ge while the contact

to MoS₂ is grounded. The p–n junction characteristics are shown in both forward and reverse bias conditions. The circled region in the forward biased characteristics shows a trend towards negative differential resistance (NDR), confirming band-to-band-tunnelling (BTBT) current (as explained in Supplementary Information S7). The measured device is shown in the inset (scale bar, 10 μm). The length and width of the device are 5.1 μm and 15 μm, respectively, and the area of overlap of MoS₂ and Ge is 54.6 μm².

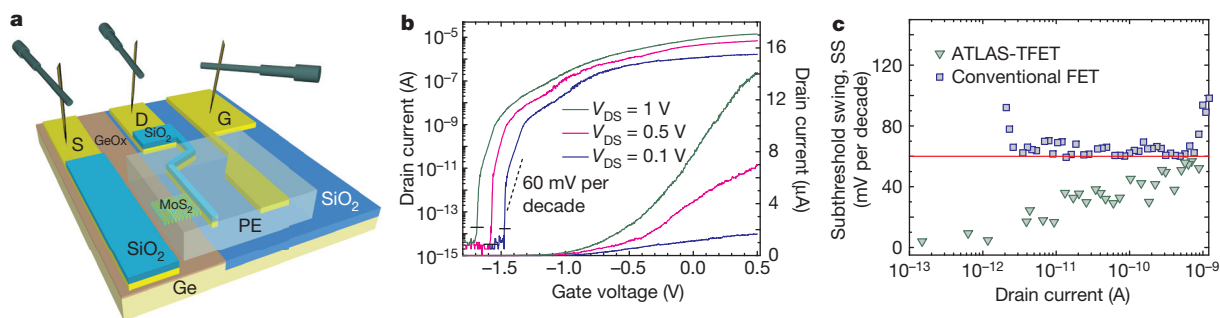


Figure 4 | Room-temperature electrical characteristics of an ATLAS-TFET. **a**, Schematic diagram showing the probing configuration for measurement of the characteristics of the ATLAS-TFET. The source and drain electrodes are covered with SiO₂ to prevent these electrodes from influencing the polymer electrolyte as well as to reduce leakage between these electrodes and the gate. **b**, Drain current as a function of gate voltage for three different drain voltages of 0.1 V, 0.5 V and 1 V. The ‘cross point’, where the drain current become similar to the gate leakage current (gate leakage characteristics are shown in Supplementary Information S9), is shown by a short black dash on each curve for a particular V_{DS} (note logarithmic scale). Below these cross points, current from the gate flows into the source and drain terminals and contaminates the source–drain current. In our case, current from the gate goes mainly to the source owing to the higher overlap between Ge and the polymer complex gate. The cross points occur at or below 3×10^{-14} A for all the V_{DS} values plotted. We derived SS over the current range above 10^{-13} A (specifically

10^{-13} – 10^{-9} A), which is well above the cross points and hence the SS is not contaminated by the gate leakage. The dashed black line indicates a slope of 60 mV per decade. Although SS may appear steeper than 60 mV per decade only for $V_{DS} = 0.5$ V, in the current range between 2 pA and 1 nA, it is also below 60 mV per decade for V_{DS} values of 0.1 V and 1 V. The average SS in the current range between 2 pA and 1 nA for a V_{DS} of 0.1 V and 1 V is 52.6 mV per decade and 46.4 mV per decade, respectively. The device measured is the same as in Fig. 3b. The shift in the curves to the left with increasing V_{DS} is due to the higher drain capacitance, which is a general feature of all TFETs (Supplementary Information S12). Note the two sets of drain current curves in **b**, one on a log scale (left set of curves and y axis) and the other on a linear scale (right set of curves and y axis). **c**, SS as a function of drain current for an ATLAS-TFET (green triangles) as well as a conventional MOSFET (blue squares) at $V_{DS} = 0.5$ V. The red line demarcates the fundamental lower limit of SS of conventional FETs.

MoS₂; this CFET exhibited an SS of >60 mV per decade, as expected (Supplementary Information S19).

Our ATLAS-TFET can potentially address both the scalability and energy-efficiency requirements of nanoscale FETs, and can guide development of next-generation ultra-low-power integrated electronics and ultra-sensitive sensors.

Received 12 February; accepted 29 July 2015.

- Novoselov, K. S. *et al.* Two-dimensional atomic crystals. *Proc. Natl Acad. Sci. USA* **102**, 10451–10453 (2005).
- Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS₂ transistors. *Nature Nanotechnol.* **6**, 147–150 (2011).
- Lee, C.-H. *et al.* Atomically thin p–n junctions with van der Waals heterointerfaces. *Nature Nanotechnol.* **9**, 676–681 (2014).
- Mak, K. F., Lee, C., Hone, J., Shan, J. & Heinz, T. F. Atomically thin MoS₂: a new direct-gap semiconductor. *Phys. Rev. Lett.* **105**, 136805 (2010).
- Wang, Q. H., Kalantar-Zadeh, K., Kis, A., Coleman, J. N. & Strano, M. S. Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. *Nature Nanotechnol.* **7**, 699–712 (2012).
- Fang, H. *et al.* High performance single layered WSe₂ p-FETs with chemically doped contacts. *Nano Lett.* **12**, 3788 (2012).
- Liu, W. *et al.* Role of metal contacts in designing high-performance monolayer n-type WSe₂ field effect transistors. *Nano Lett.* **13**, 1983–1990 (2013).
- Lundstrom, M. S. The MOSFET revisited: device physics and modeling at the nanoscale. In *IEEE International SOI Conference Proceedings* 17–19 (IEEE, 2006).
- Sakurai, T. Perspectives of low-power VLSI's. *IEICE Trans. Electron.* **E87-C**, 429–436 (2004).
- Quinn, J. J. & Kawamoto, G. Subband spectroscopy by surface channel tunneling. *Surf. Sci.* **73**, 190–196 (1978).
- Baba, T. Proposal for surface tunnel transistors. *Jpn. J. Appl. Phys.* **31**, L455–L457 (1992).
- Bhuvalka, K. K. *et al.* Vertical tunnel field-effect transistor. *IEEE Trans. Electron. Dev.* **51**, 279–282 (2004).
- Zhang, Q., Zhao, W., Member, S. & Seabaugh, A. Low-subthreshold-swing tunnel transistors. *IEEE Electron Device Lett.* **27**, 297–300 (2006).
- Khatami, Y. & Banerjee, K. Steep subthreshold slope n- and p-type tunnel-FET devices for low-power and energy-efficient digital circuits. *IEEE Trans. Electron. Dev.* **56**, 2752–2761 (2009).
- Ionescu, A. M. & Riel, H. Tunnel field-effect transistors as energy-efficient electronic switches. *Nature* **479**, 329–337 (2011).
- Datta, S., Liu, H. & Narayanan, V. Tunnel FET technology: a reliability perspective. *Microelectron. Reliab.* **54**, 861–874 (2014).
- International Technology Roadmap for Semiconductors. <http://www.itrs.net/ITRS%201999-2014%20Mtg%20Presentations%20&%20Links/2013ITRS/Summary2013.htm> (2013).
- Sarkar, D. & Banerjee, K. Proposal for tunnel-field-effect-transistor as ultra-sensitive and label-free biosensors. *Appl. Phys. Lett.* **100**, 143108 (2012).
- Rodgers, P. Biomolecular turn-ons. *Nature Nanotechnol.* **7**, 275 (2012).
- Sarkar, D., Gossner, H., Hansch, W. & Banerjee, K. Tunnel-field-effect-transistor based gas-sensor: introducing gas detection with a quantum-mechanical transducer. *Appl. Phys. Lett.* **102**, 023110 (2013).
- Sarkar, D. *et al.* MoS₂ field-effect transistor for next-generation label-free biosensors. *ACS Nano* **8**, 3992–4003 (2014).
- Sze, S. M. & Ng, K. *Physics of Semiconductor Devices* 3rd edn (Wiley, 2008).
- Sarkar, D., Krall, M. & Banerjee, K. Electron-hole duality during band-to-band tunneling process in graphene-nanoribbon tunnel-field-effect-transistors. *Appl. Phys. Lett.* **97**, 263109 (2010).
- Das, S., Prakash, A., Salazar, R. & Appenzeller, J. Toward low-power electronics: tunneling phenomena in transition metal dichalcogenides. *ACS Nano* **8**, 1681–1689 (2014).
- Roy, T. *et al.* Dual-gated MoS₂/WSe₂ van der Waals tunnel diodes and transistors. *ACS Nano* **9**, 2071–2079 (2015).
- Britnell, L. *et al.* Field-effect tunneling transistor based on vertical graphene heterostructures. *Science* **335**, 947–950 (2012).
- Georgiou, T. *et al.* Vertical field-effect transistor based on graphene-WSe₂ heterostructures for flexible and transparent electronics. *Nature Nanotechnol.* **8**, 100–103 (2013).
- Tomioka, K., Yoshimura, M. & Fukui, T. Steep-slope tunnel field-effect transistors using III–V nanowire/Si heterojunction. In *IEEE Symposium on VLSI Technology* 47–48 (IEEE, 2012).
- Liu, W. *et al.* High-performance few-layer-MoS₂ field-effect-transistor with record low contact-resistance. In *IEEE International Electron Devices Meeting* 499–502 (IEEE, 2013).
- Lin, M.-W. *et al.* Mobility enhancement and highly efficient gating of monolayer MoS₂ transistors with polymer electrolyte. *J. Phys. D* **45**, 345102 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Air Force Office of Scientific Research (grant FA9550-14-1-0268) and the US NSF (grant CCF-1162633). Y.G. was supported by the Army Research Office (MURI grant W911NF-11-1-0362). All process steps for device fabrication were carried out using the Nanostructure Cleanroom Facility at the California NanoSystems Institute and the Nanofabrication Facilities at UCSB—part of the National Nanotechnology Infrastructure Network. We made extensive use of the MRL Central Facilities at UCSB, which are supported by the MRSEC Program of the NSF (award no. DMR 1121053), a member of the NSF-funded Materials Research Facilities Network (<http://www.mrfn.org>).

Author Contributions K.B. initiated, planned and led the research. D.S. designed and fabricated the devices, collected and analysed the data, with input from X.X., W.L., W.C. and J.K. X.X. and W.L. performed Raman analysis. W.C. collected and analysed data on previously demonstrated tunnel-FET devices. Y.G. and P.M.A. synthesized large-area bilayered MoS₂ samples. S.K. performed TEM analysis. D.S. and K.B. wrote the main Letter and the Supplementary Information with input from all other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.B. (kaustav@ece.ucsb.edu).

Binding of dinitrogen to an iron–sulfur–carbon site

Ilija Čorić¹, Brandon Q. Mercado¹, Eckhard Bill², David J. Vinyard¹ & Patrick L. Holland¹

Nitrogenases are the enzymes by which certain microorganisms convert atmospheric dinitrogen (N_2) to ammonia, thereby providing essential nitrogen atoms for higher organisms. The most common nitrogenases reduce atmospheric N_2 at the FeMo cofactor, a sulfur-rich iron–molybdenum cluster (FeMoCo)^{1–5}. The central iron sites that are coordinated to sulfur and carbon atoms in FeMoCo have been proposed to be the substrate binding sites, on the basis of kinetic and spectroscopic studies^{5–7}. In the resting state, the central iron sites each have bonds to three sulfur atoms and one carbon atom. Addition of electrons to the resting state causes the FeMoCo to react with N_2 , but the geometry and bonding environment of N_2 -bound species remain unknown⁵. Here we describe a synthetic complex with a sulfur-rich coordination sphere that, upon reduction, breaks an Fe–S bond and binds N_2 . The product is the first synthetic Fe– N_2 complex in which iron has bonds to sulfur and carbon atoms, providing a model for N_2 coordination in the FeMoCo. Our results demonstrate that breaking an Fe–S bond is a chemically reasonable route to N_2 binding in the FeMoCo, and show structural and spectroscopic details for weakened N_2 on a sulfur-rich iron site.

The FeMoCo does not bind N_2 in the resting state of nitrogenase (structure I, Fig. 1a); instead N_2 binds after the addition of electrons, and the structure of the intermediate is unknown^{5,6,8}. Most researchers have hypothesized that N_2 binding to FeMoCo takes place at an iron centre with three sulfur ligands following Fe–C bond elongation or dissociation (I to II, Fig. 1a)^{9–15}. Here, we focus on an alternative

hypothesis where one of the Fe–S bonds at the active site is broken upon reduction/protonation to expose the N_2 binding site (I to III, Fig. 1a)^{16,17}. N_2 would thus bind at a pseudotetrahedral S₃C-bound iron site. The feasibility of Fe–S bond cleavage in FeMoCo is experimentally supported by the loss of this sulfur atom in the structure of CO-inhibited nitrogenase⁷, and by the observation of Fe–S cleavage upon protonation in smaller Fe–S clusters^{18,19}. Other N_2 binding hypotheses include side-on binding, bridging, and *endo* coordination in which N_2 is positioned close to three additional iron atoms and opposite to a sulfur atom (IV, Fig. 1a)^{5,11,17}.

Fe– N_2 complexes supported solely by sulfur, or by sulfur and carbon ligands, are probable N_2 -bound species in the nitrogenase catalytic cycle, but they are experimentally unprecedented. Though chemists have prepared complex Fe–S clusters inspired by the multimetallic structure of FeMoCo, N_2 does not bind to any known synthetic Fe–S cluster²⁰. A number of well defined iron complexes with B, N, C, and P supporting ligands are known to activate N_2 (refs 21–24), and Peters has established systems capable of performing catalytic reduction of N_2 to ammonia^{14,15,21}. A few Fe– N_2 complexes have thioether/thiolate donors on the same iron centre, and each is additionally supported by P- or N-donors^{25–27}. To the best of our knowledge, there are no examples of terminal N_2 complexes of any metal having immediate ligand environments similar to those in II–IV, which makes it difficult to predict the behaviour of the FeMoCo.

For this work, we designed bis(thiolate) ligand L^{2-} , which offers only sulfur- and carbon-based coordination sites (indicated by orange

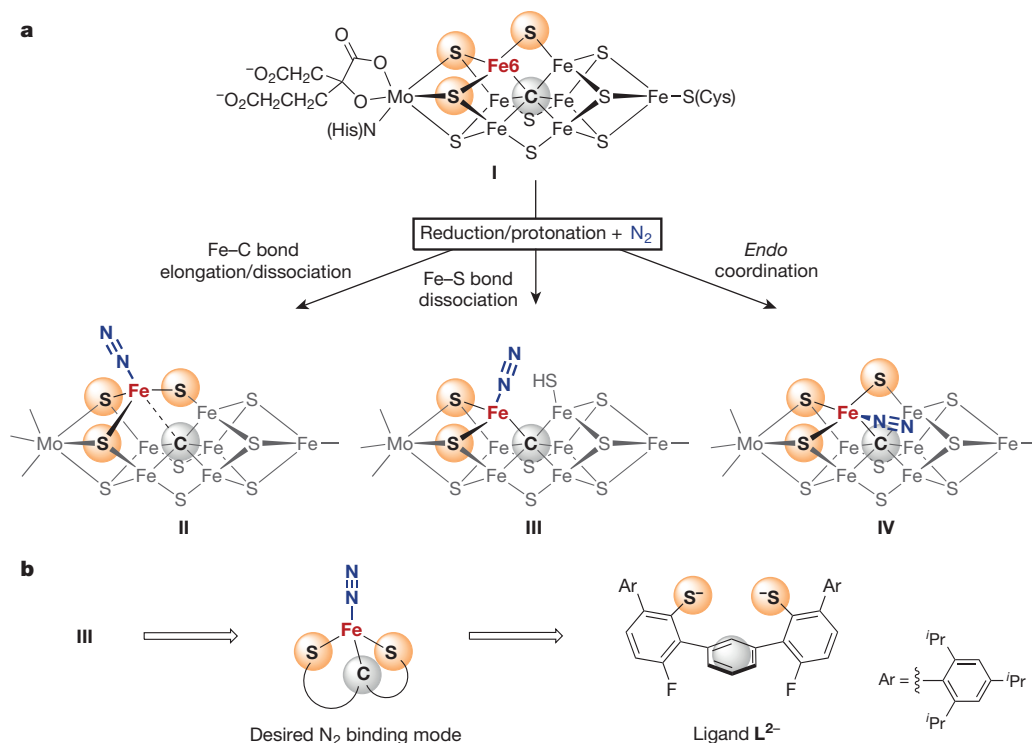


Figure 1 | N_2 binding to iron in sulfur-rich environments.

a, Schematic representations of FeMoCo and three potential N_2 -binding modes. Potentially protonated sulfur ligands are not specified. **b**, Ligand design for a synthetic sulfur–carbon site. ⁱPr = isopropyl; ^{Ar} = 2,4,6-triisopropylphenyl.

¹Department of Chemistry, Yale University, 225 Prospect Street, New Haven, Connecticut 06520, USA. ²Max Planck Institute for Chemical Energy Conversion, Stiftstrasse 34–36, D-45470 Mülheim an der Ruhr, Germany.

and grey spheres in Fig. 1b). Our approach was guided by the proposed binding mode **III** in Fig. 1a, which requires the presence of two coordinating sulfur atoms. These are provided by two chelating arylthiolate donors with bulky 2,4,6-triisopropylphenyl groups shielding the sulfur sites. A central aromatic ring connects the two arylthiolate arms and additionally provides potential carbon-based attachment sites²⁸. Although carbide is electronically different to the arene ring in L^{2-} , either could provide flexible bonding for the stabilization of various intermediates during ammonia production^{14,15}.

Iron(II) ions were installed in the ligand framework by treating **LH**₂ with iron(II) bis(bis(trimethylsilyl)amide) in tetrahydrofuran (THF), which yielded the bright yellow, high spin iron(II) complex **1** (Fig. 2a). Its crystal structure reveals that it is four-coordinate, and that all Fe–C distances are at least 2.59 Å (Fig. 2b). Reduction of **1** to iron(I) with potassium graphite (KC₈) forms the brownish-yellow product **2**, with close Fe–C distances (2.04–2.15 Å) indicating η^6 -binding of the central arene ring (Fig. 2a and c). Comparison of the molecular structures of **1** and **2** reveals that rotation of the arylthiolate arms enables the central aryl ring to move closer to the iron atom. Compound **2** has a rhombic EPR (electron paramagnetic resonance) spectrum with principal *g* values of 2.180, 2.020, and 1.989, and a solution magnetic moment of 2.1 μ_B , (where μ_B is the Bohr magneton) both of which indicate a low-spin (*S* = 1/2) iron(I) centre.

Encouraged by the ability of the ligand to stabilize low-valent iron sites, we further reduced the iron site to the iron(0) oxidation state. Reduction of a brownish-yellow solution of **2** with one equivalent of KC₈ under one atmosphere of N₂ at –70 °C resulted in an immediate colour change to deep red. After addition of 18-crown-6 to sequester potassium cations, dark red-brown crystals of **3** grew at –40 °C. X-ray diffraction analysis shows that **3** is [LFeN₂][K(18-crown-6)(THF)]₂

(Figs 2a and 3a). In **3**, N₂ is bound as a terminal ligand at a pseudo-tetrahedral iron(0) site, which is further bound to two sulfur atoms and the arene of the supporting ligand. The closest Fe–C distance in **3** is 2.04 Å, and there is a second carbon atom within bonding distance (Fe–C = 2.24 Å), indicating asymmetric η^2 coordination of the arene. The potassium cations do not bind to the N₂ ligand.

The new N₂ complex **3** provides a structural model of the pseudo-tetrahedral S₂C-supported N₂-binding mode **III** proposed for FeMoco (Fig. 1a). It is compared to the experimental structures of resting-state FeMoco and CO-inhibited FeMoco in Fig. 3b^{2,7}. In the fourth coordination site that has labile S and CO ligands in nitrogenase structures⁷, **3** contains an N₂ ligand. The Fe–S bond distances in **3** (2.32–2.35 Å) are somewhat longer than the Fe–S bonds in resting-state FeMoco (2.25–2.27 Å), owing to either the lesser negative charge of the thiolate or the greater steric hindrance. Remarkably, the Fe–C distance in **3** at 2.04 Å is very close to the Fe6–carbide distance of 2.01 Å in FeMoco structures. Overall, the relatively simple ligand L^{2-} is capable of arranging appropriate atoms around iron and imparting a geometry that resembles the likely active iron site in FeMoco structures. However, the electronic structure of the iron(0) complex **3** may be different to that of the iron site in the N₂-binding form of the FeMoco (for which the structure and iron oxidation state are unknown).

Next, we designed a compound (**5**) intended to test the idea that Fe–S bond dissociation could provide a coordination site for N₂ binding (**I** to **III** in Fig. 1a). The bis(thiolate) complex **1** reacted with thiolate **4** to give the iron tris(thiolate) complex **5** (Fig. 2a). This orange high-spin iron(II) complex contains three S ligands, like Fe6 in the FeMoco resting state (**I** in Fig. 1a). The interaction of iron with the central arene ring is weak, with the closest Fe–C distance at 2.48 Å (Fig. 2d). Thus we

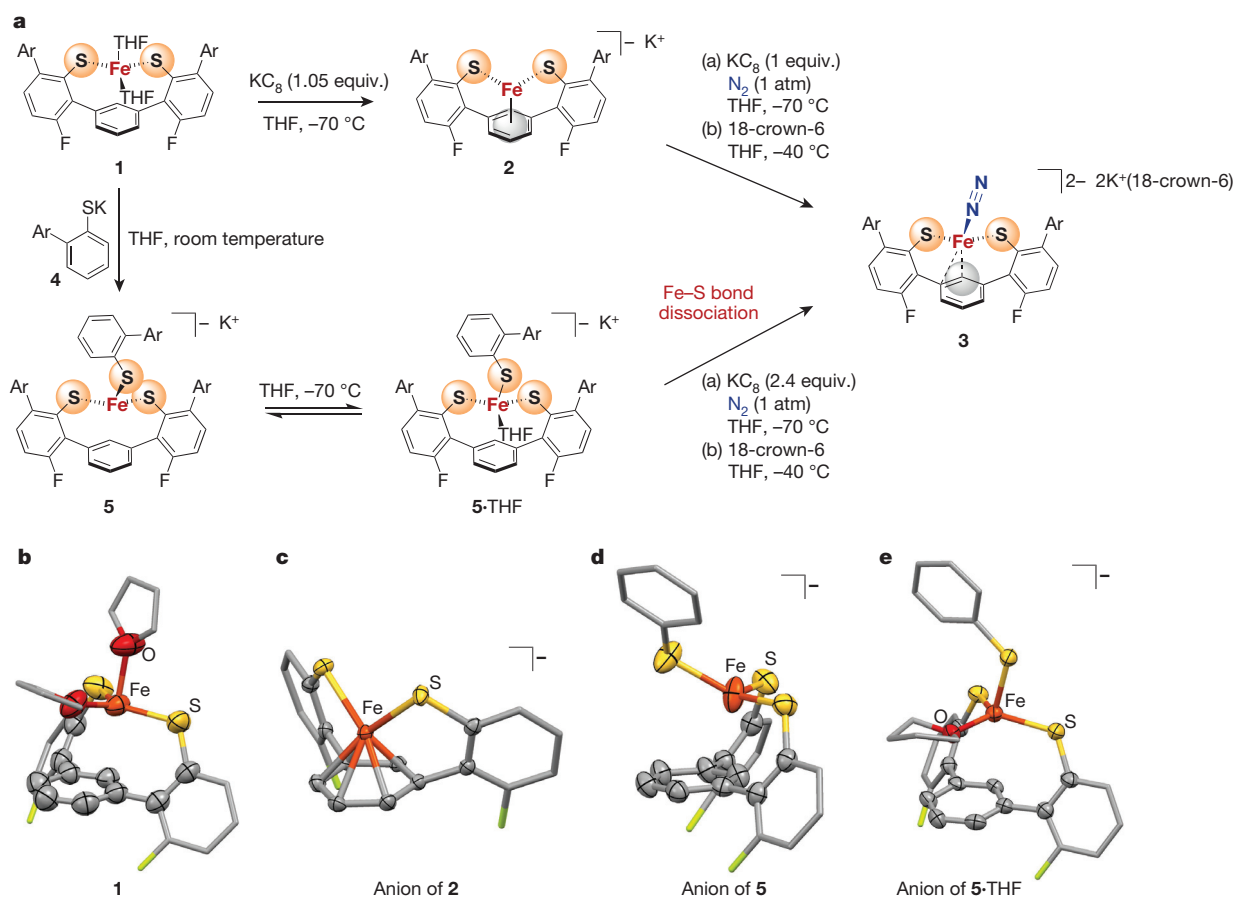


Figure 2 | N₂ binding at an iron-sulfur-carbon site through Fe–S bond cleavage. **a**, Reactions of synthetic Fe–S sites leading to N₂ binding. The bottom pathway shows Fe–S cleavage with N₂ binding. **b–e**, Molecular structures of the

synthetic mononuclear Fe–S sites presented here. Hydrogen atoms and 'Ar' groups are omitted for clarity.

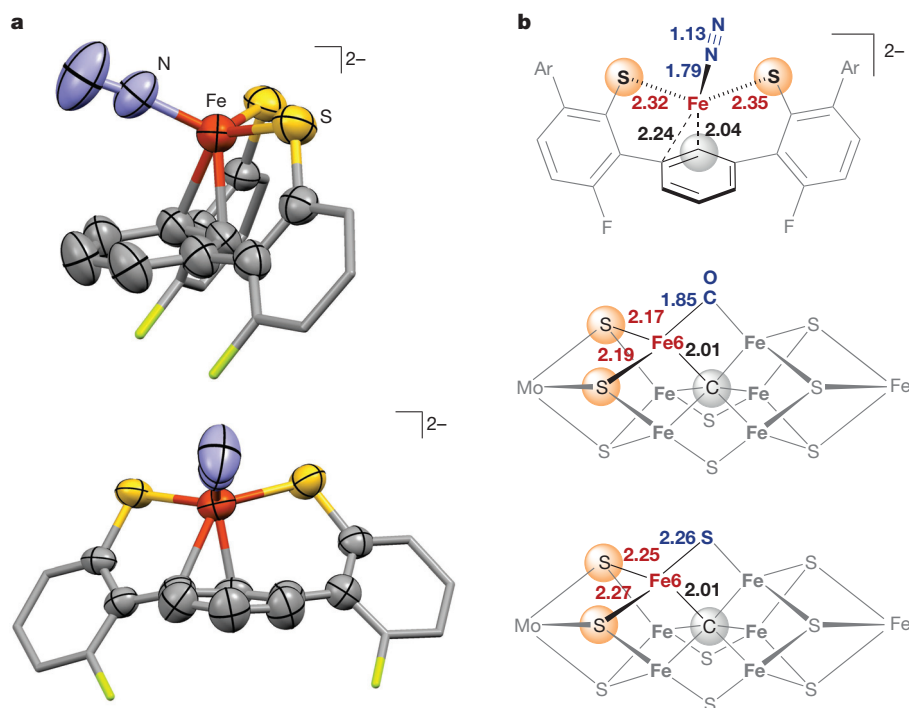


Figure 3 | Fe–N₂ complex supported by sulfur and carbon ligands. **a**, Two views of the molecular structure of the anionic part of **3**. Hydrogen atoms and ‘Ar’ groups are omitted. **b**, Comparison of geometric parameters with CO-inhibited FeMoco⁷ and resting-state FeMoco³. All distances are reported in ångströms.

view this site as three-coordinate and unsaturated, which is supported by the reversible binding of one THF molecule at low temperature (Fig. 2e and Supplementary Information show the X-ray crystal structure of **5**·THF and temperature-dependent ultraviolet–visible spectra).

The tris(thiolate) iron(II) site in this equilibrium mixture of **5** and **5**·THF was reduced to the iron(0) oxidation level with just over two equivalents of K⁺C₈, under conditions otherwise equivalent to those used for reduction of the iron(I) bis(thiolate) complex **2** (Fig. 2a). This yielded the same N₂ complex **3** described above, and 1.0 equivalent of free thiolate was produced. Reduction thus causes an Fe–S bond to break concomitant with N₂ binding, as in the proposed pathway for N₂ binding to FeMoco in Fig. 1a (I to III). We note that tris(thiolate) **5** contains all the nearby atoms to support alternative binding modes II and IV in Fig. 1a, but Fe–S dissociation takes place instead.

We return to describe the further characterization of **3**, which gives insight into potential properties of N₂ after binding at FeMoco. Although complex **3** is very thermally sensitive, it was possible to isolate pure samples of **3** in >80% yield from reduction of **5** at low temperature and washing the crystals with cold butane at –70 °C. Analysis of these crystals by Mössbauer spectroscopy confirms the presence of a single iron species. Infrared spectroscopy analysis of single crystals of **3** revealed a strong N–N stretching band at 1,880 cm^{–1}. These frequencies are the lowest observed for any Fe–N₂ complex with a terminal N₂ ligand²³, which shows that the thiolates are powerful electron donors that enable substantial backbonding from the Fe *d* orbitals into the N₂ π* orbitals. The N₂ ligand in **3** exchanges with free ¹⁵N₂ (giving an ¹⁵N–¹⁵N stretching band at 1,813 cm^{–1}) at –70 °C in the solid state. Samples of **3** kept at room temperature for a few hours lack the N₂ stretching vibration, further demonstrating the lability of N₂. The lability suggests that the Fe–N₂ interaction, though strong as judged by infrared spectroscopy, may be replaced with tighter binding to the arene ring.

Compound **3** has a high-spin (*S* = 1) electronic configuration, as determined by SQUID (superconducting quantum interference device) magnetometry on a crystalline sample. This experimental observation was confirmed with density functional theory calculations on a truncated model of **3**. Optimization with *S* = 1 gave a model close to the experimental geometry, but optimization with *S* = 0 gave substantially different bond lengths and angles, and a Gibbs free energy

(Δ*G*^o) that was higher by 37 kJ mol^{–1} (see Supplementary Information). High-spin iron(0) N₂ complexes are rare, and have been seen mainly in cases where high symmetry makes the frontier orbitals nearly degenerate^{29,30}. To our knowledge, **3** is the first high-spin iron complex that contains both S and N₂ ligands^{25,26}, and shows that high-spin iron (as expected in the weak-field sulfur-dominated environment of iron atoms in the FeMoco) can substantially activate N₂.

The preparation of an Fe–N₂ complex with a sulfur-rich environment provides structural and spectroscopic precedents for FeMoco–N₂ binding, and also gives insight into the nitrogenase mechanism. Reduction of complex **5** breaks an Fe–S bond as in the hypothetical conversion of I to III in the FeMoco (Fig. 1a), and binds N₂ in a form where the N–N bond is weakened. In this way, the results support the idea that the sulfur-rich iron site in the FeMoco is particularly well suited for N₂ activation, and that Fe–S bonds can be easily broken upon reduction to allow binding of N₂.

Received 27 May; accepted 24 July 2015.

Published online 23 September 2015.

1. Einsle, O. *et al.* Nitrogenase MoFe-protein at 1.16 Å resolution: a central ligand in the FeMo-cofactor. *Science* **297**, 1696–1700 (2002).
2. Spatzal, T. *et al.* Evidence for interstitial carbon in nitrogenase FeMo cofactor. *Science* **334**, 940 (2011).
3. Lancaster, K. M. *et al.* X-ray emission spectroscopy evidences a central carbon in the nitrogenase iron-molybdenum cofactor. *Science* **334**, 974–977 (2011).
4. Wiig, J. A., Hu, Y., Lee, C. C. & Ribbe, M. W. Radical SAM-dependent carbon insertion into the nitrogenase M-cluster. *Science* **337**, 1672–1675 (2012).
5. Hoffman, B. M., Lukoyanov, D., Yang, Z.-Y., Dean, D. R. & Seefeldt, L. C. Mechanism of nitrogen fixation by nitrogenase: the next stage. *Chem. Rev.* **114**, 4041–4062 (2014).
6. Seefeldt, L. C., Hoffman, B. M. & Dean, D. R. Mechanism of Mo-dependent nitrogenase. *Annu. Rev. Biochem.* **78**, 701–722 (2009).
7. Spatzal, T., Perez, K. A., Einsle, O., Howard, J. B. & Rees, D. C. Ligand binding to the FeMo-cofactor: structures of CO-bound and reactivated nitrogenase. *Science* **345**, 1620–1623 (2014).
8. Yandulov, D. V. & Schrock, R. R. Catalytic reduction of dinitrogen to ammonia at a single molybdenum center. *Science* **301**, 76–78 (2003).
9. Holland, P. L. Low-coordinate iron complexes as synthetic models of nitrogenase. *Can. J. Chem.* **83**, 296–301 (2005).
10. MacBeth, C. E., Harkins, S. B. & Peters, J. C. Synthesis and characterization of cationic iron complexes supported by the neutral ligands NP⁺Pr₃, NAr⁺Pr₃, and NS⁺Bu₃. *Can. J. Chem.* **83**, 332–340 (2005).
11. Dance, I. Ramifications of C-centering rather than N-centering of the active site FeMo-co of the enzyme nitrogenase. *Dalton Trans.* **41**, 4859–4865 (2012).

12. Hinnemann, B. & Nørskov, J. K. Chemical activity of the nitrogenase FeMo cofactor with a central nitrogen ligand: density functional study. *J. Am. Chem. Soc.* **126**, 3920–3927 (2004).
13. George, S. J. *et al.* EXAFS and NRVs reveal a conformational distortion of the FeMo-cofactor in the MoFe nitrogenase propargyl alcohol complex. *J. Inorg. Biochem.* **112**, 85–92 (2012).
14. Creutz, S. E. & Peters, J. C. Catalytic reduction of N₂ to NH₃ by an Fe–N₂ complex featuring a C-atom anchor. *J. Am. Chem. Soc.* **136**, 1105–1115 (2014).
15. Anderson, J. S., Rittle, J. & Peters, J. C. Catalytic conversion of nitrogen to ammonia by an iron model complex. *Nature* **501**, 84–87 (2013).
16. Kästner, J. & Blöchl, P. E. Ammonia production at the FeMo cofactor of nitrogenase: results from density functional theory. *J. Am. Chem. Soc.* **129**, 2998–3006 (2007).
17. Schimpl, J., Petrilli, H. M. & Blöchl, P. E. Nitrogen binding to the FeMo-cofactor of nitrogenase. *J. Am. Chem. Soc.* **125**, 15772–15778 (2003).
18. Alwaaly, A., Dance, I. & Henderson, R. A. Unexpected explanation for the enigmatic acid-catalysed reactivity of [Fe₄S₄X₄]^{2–} clusters. *Chem. Commun.* **50**, 4799–4802 (2014).
19. Saouma, C. T., Morris, W. D., Darcy, J. W. & Mayer, J. M. Protonation and proton-coupled electron transfer at S-ligated [4Fe-4S] clusters. *Chem. Eur. J.* **21**, 9256–9260 (2015).
20. Lee, S. C., Lo, W. & Holm, R. H. Developments in the biomimetic chemistry of cubane-type and higher nuclearity iron–sulfur clusters. *Chem. Rev.* **114**, 3579–3600 (2014).
21. Ung, G. & Peters, J. C. Low-temperature N₂ binding to two-coordinate L₂Fe⁰ enables reductive trapping of L₂FeN₂[–] and NH₃ generation. *Angew. Chem. Int. Ed.* **54**, 532–535 (2015).
22. Rodríguez, M. M., Bill, E., Brennessel, W. W. & Holland, P. L. N₂ reduction and hydrogenation to ammonia by a molecular iron–potassium complex. *Science* **334**, 780–783 (2011).
23. Hazari, N. Homogeneous iron complexes for the conversion of dinitrogen into ammonia and hydrazine. *Chem. Soc. Rev.* **39**, 4044–4056 (2010).
24. Danopoulos, A. A., Wright, J. A. & Motherwell, W. B. Molecular N₂ complexes of iron stabilised by N-heterocyclic ‘pincer’ dicarbene ligands. *Chem. Commun.* 784–786 (2005).
25. Takaoka, A., Mankad, N. P. & Peters, J. C. Dinitrogen complexes of sulfur-ligated iron. *J. Am. Chem. Soc.* **133**, 8440–8443 (2011).
26. Bart, S. C., Lobkovsky, E., Bill, E., Wieghardt, K. & Chirik, P. J. Neutral-ligand complexes of bis(imino)pyridine iron: synthesis, structure, and spectroscopy. *Inorg. Chem.* **46**, 7055–7063 (2007).
27. Creutz, S. E. & Peters, J. C. Diiron bridged-thiolate complexes that bind N₂ at the Fe^{II}Fe^{II}, Fe^{II}Fe^I, and Fe^IFe^I redox states. *J. Am. Chem. Soc.* **137**, 7310–7313 (2015).
28. Ellison, J. J., Ruhlandt-Senge, K. & Power, P. P. Synthesis and characterization of thiolato complexes with two-coordinate iron(II). *Angew. Chem. Int. Edn Engl.* **33**, 1178–1180 (1994).
29. Suess, D. L. M. & Peters, J. C. H–H and Si–H bond addition to Fe≡NNR₂ intermediates derived from N₂. *J. Am. Chem. Soc.* **135**, 4938–4941 (2013).
30. Moret, M.-E. & Peters, J. C. Terminal iron dinitrogen and iron imide complexes supported by a tris(phosphino)borane ligand. *Angew. Chem. Int. Ed.* **50**, 2063–2067 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the National Institutes of Health (GM065313 to P.L.H.) and the Max Planck Society (E.B.). We thank A. Göbels for measurement of SQUID data and G. Brudvig for the use of an EPR spectrometer. Elemental analysis data were from the CENTC Elemental Analysis Facility at the University of Rochester, funded by the NSF (CHE-0650456), and we thank W. Brennessel for collecting these data. This work was supported in part by the facilities and staff of the Yale High Performance Computing Center, which was partially funded by the NSF (CNS 08-21132). We thank J. Mayer, N. Hazari, S. Bonyhady, N. Arnet, and C. MacLeod for constructive criticism on the manuscript.

Author Contributions I.C. designed the iron–sulfur–carbon system for N₂ binding, performed the laboratory experiments, and analysed data. B.Q.M. collected and interpreted crystallographic data. E.B. interpreted solid-state (SQUID) magnetic data. D.J.V. collected and fitted EPR data. P.L.H. supervised the research, and I.C. and P.L.H. wrote the manuscript.

Author Information X-ray crystallographic data have been deposited in the Cambridge Crystallographic Data Centre (<http://www.ccdc.cam.ac.uk/>) with deposition numbers CCDC1402555–CCDC1402559. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.L.H. (patrick.holland@yale.edu).

Observed latitudinal variations in erosion as a function of glacier dynamics

Michèle Koppes¹, Bernard Hallet², Eric Rignot^{3,4}, Jérémie Mouginot³, Julia Smith Wellner⁵ & Katherine Boldt⁶

Glacial erosion is fundamental to our understanding of the role of Cenozoic-era climate change in the development of topography worldwide, yet the factors that control the rate of erosion by ice remain poorly understood. In many tectonically active mountain ranges, glaciers have been inferred to be highly erosive, and conditions of glaciation are used to explain both the marked relief typical of alpine settings and the limit on mountain heights above the snowline, that is, the glacial buzzsaw¹. In other high-latitude regions, glacial erosion is presumed to be minimal, where a mantle of cold ice effectively protects landscapes from erosion^{2–4}. Glacial erosion rates are expected to increase with decreasing latitude, owing to the climatic control on basal temperature and the production of meltwater, which promotes glacial sliding, erosion and sediment transfer. This relationship between climate, glacier dynamics and erosion rate is the focus of recent numerical modelling^{5–8}, yet it is qualitative and lacks an empirical database. Here we present a comprehensive data set that permits explicit examination of the factors controlling glacier erosion across climatic regimes. We report contemporary ice fluxes, sliding speeds and erosion rates inferred from sediment yields from 15 outlet glaciers spanning 19 degrees of latitude from Patagonia to the Antarctic Peninsula. Although this broad region has a relatively uniform tectonic and geologic history, the thermal regimes of its glaciers range from temperate to polar. We find that basin-averaged erosion rates vary by three orders of magnitude over this latitudinal transect. Our findings imply that climate and the glacier thermal regime control erosion rates more than do extent of ice cover, ice flux or sliding speeds.

Our ability to assess how glacial erosion shapes mountain ranges and reflects climate or tectonic variability is limited by a dearth of information about what controls the rate of glacial erosion, today and in the past. Maximum erosion rates can surpass those of fluvial erosion by up to an order of magnitude^{9,10}, but the few available data sets indicative of rapid glacial erosion are predominantly from massive, fast-moving, temperate tidewater glaciers^{9–11}. In polar regions and many high-altitude alpine settings, glacial erosion is markedly slower^{12,13}. The available data report a wide range of erosion rates from individual ice masses over varying timescales^{9,10,14}, but the cause of this wide range has not been addressed, primarily because of the lack of substantive complementary glaciological data on the ice masses responsible for the erosion.

Recent numerical models have focused on processes that produce glacial landscapes^{5–8,15}. Central to these models is a simple index that relates erosion rate to ice dynamics. Most models assume that erosion rates are proportional to the sliding velocity at the bed^{5–8,15} or the integrated ice discharge¹⁶, and that they reach a peak at the equilibrium-line altitude (ELA). Theory strongly ties the rate of glacial erosion by quarrying and abrasion to the rate of sliding and the effective pressure at the bed, which is controlled by climate through the glacial

thermal regime, ice flux and the amount of meltwater produced^{17,18}. The large meltwater discharge typical of temperate glacial systems evacuates large amounts of debris from under the ice, resulting in massive sediment accumulation at the terminus^{19,20}. In colder climates, glacial erosion is expected to decrease progressively as surface melting decreases, because little or no water reaches the bed to facilitate glacier sliding and flush out any sediment generated from erosion^{21,22}.

Although treatments of glacier dynamics in numerical models have a firm theoretical basis and have become increasingly sophisticated^{5,6}, the parameters that relate erosion to basal sliding remain poorly constrained^{6,7,18}. Most models use a bedrock ‘erosion rule’ of the form $E = K_g u_s^n$, where u_s is the glacier sliding speed, K_g is a constant representing bedrock erosion susceptibility (varying between 10^{-4} and 10^{-6}), and n is a constant that is normally assumed to be one^{5–8,15}. In most cases, the two constants are based on a single empirical study in which both the sediment yield and ice motion were measured at Variegated Glacier, Alaska²³ over a period of a few years that included a major glacier surge. The data we present clearly suggest that the scaling between erosion and sliding rates is strongly affected by the glacier thermal regime.

In an effort to fill this data gap and provide a quantitative test of long-held assumptions, we examine explicitly the factors controlling modern glacier erosion rates across a wide range of climatic regimes. We chose 15 tidewater-outlet glaciers extending from northern Patagonia to the western Antarctic Peninsula (Fig. 1), an area that spans almost 20° of latitude and whose mean annual air temperature varies by 14 °C. This area is an ideal natural laboratory for our purposes: it covers a broad region with a relatively uniform recent tectonic history and bedrock lithologies; it contains a climatically diverse range of glacier thermal regimes that vary from temperate to polar and a transect of glaciers with similar catchment hypsometries; and it is a region where the fjords constitute accessible, natural sediment traps for the products of erosion from the watersheds over the past century. To our knowledge, this is the only transect combining quantitative measurements of both glacier dynamics and erosion rates assembled so far.

Prior studies of sediment accumulation in the region have documented a substantial decrease in sedimentation in the fjords from north to south and west to east^{24–26}, which has been inferred to reflect climate-driven differences in glacier dynamics and meltwater. Here we take a more quantitative approach and focus on the past 50–100 years, for which considerable data exist on both the glaciers and sediment yields. For each glacier catchment, we estimate the contemporary sediment yield using two complementary approaches, as required by the large range of sedimentation rates and the time span of interest. First, we calculate the sediment volume in recently deglaciated fjords from acoustic reflection profiles, repeat bathymetry and the history of glacial retreat¹¹ (for all but one of the glaciers in Patagonia). Second, we calculate the product of the accumulation rate using published ²¹⁰Pb

¹Department of Geography, 1984 West Mall, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada. ²Department of Earth and Space Sciences and Quaternary Research Center, Box 351310, University of Washington, Seattle, Washington 98195-1310, USA. ³Department of Earth System Science, University of California, Irvine, California 92617, USA. ⁴NASA Jet Propulsion Laboratory, Pasadena, California 91109, USA. ⁵Department of Earth and Atmospheric Sciences, University of Houston, Houston, Texas 77204, USA. ⁶School of Oceanography, Box 357940, University of Washington, Seattle, Washington 98195-7940, USA.

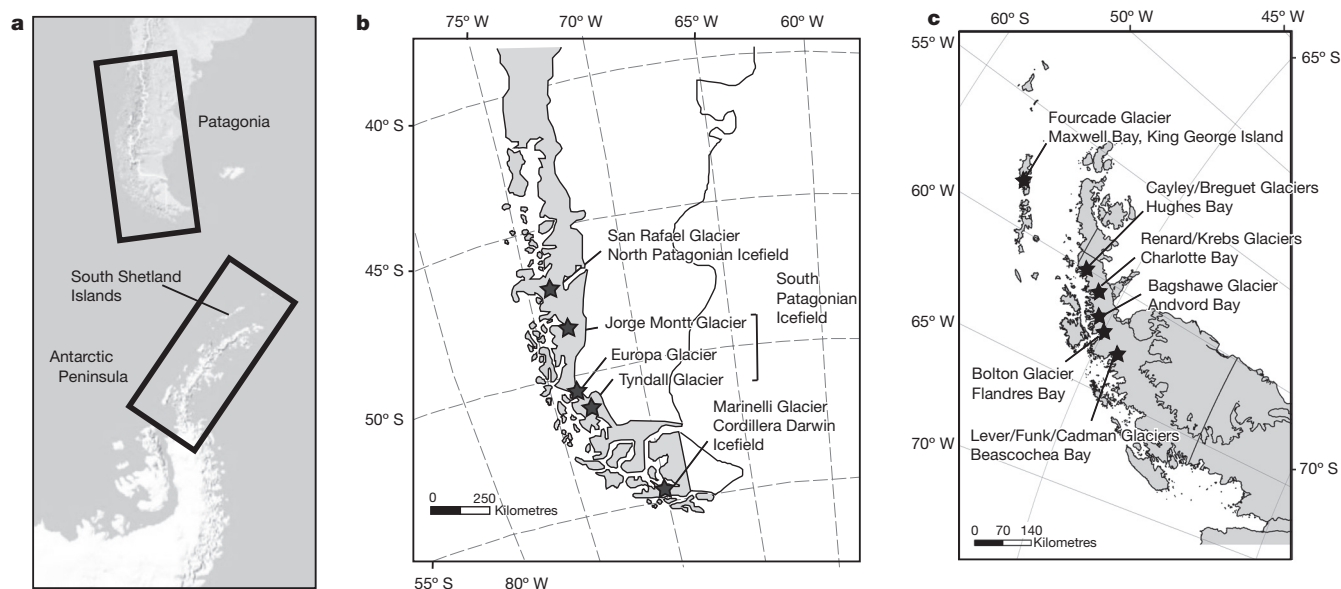


Figure 1 | Location map of study area, from northern Patagonia to central Antarctic Peninsula. **a**, Entire latitudinal transect from 46° S to 66° S. **b**, **c**, Location of glaciers and adjacent fjords in Chilean Patagonia (**b**), and the South Shetland Islands and the Antarctic Peninsula (**c**).

analyses of sediment cores^{20,27} and the spatial extent of the fjord depocentres (zones of subaqueous sediment deposition) measured from multibeam swath bathymetry. We determine the basin-averaged erosion rates over the past 50–100 years on the basis of the ratio of sediment yield to the area of the glacier-drainage basin. We then compare the erosion rates to simple indices of the contemporary dynamics of the glaciers that produced and delivered the sediment, that is, the ice flux and basal sliding velocity at the ELA. This study highlights the influence of both climate and glacier dynamics on the broad spatial pattern of contemporary glacial erosion, and leads to broader insights into the temporal variation in erosion rates seen in glaciated orogens. We take a ‘trading space for time’ approach by comparing glacier systems with similar basin characteristics but differing climatic regimes to quantify the impact of climate change on erosion rates over glacial–interglacial cycles.

The climatic conditions from the Patagonian Andes to the western Antarctic Peninsula support ice masses on mountains that rise up to 3,500 m above sea level, with glaciers terminating at sea level. In Patagonia, the ELA ranges from 700 m to 1,300 m above sea level²⁸; in the western Antarctic Peninsula, on the basis of the late summer snowlines, the ELA is just above the calving fronts²⁴. All of our study sites are north of 70° S in the ‘wet snow’ zone, where glaciers experience occasional to frequent surface melting in late summer²⁹, which suggests the potential for meltwater to access the glacier bed and for sliding and active erosion at the ice–bed interface.

On the timescale of the last 50–100 years, basin-averaged bedrock erosion rates along our 19° N–S transect vary by three orders of magnitude, largely as a function of temperature (Fig. 2). For each glacier, the erosion rate and corresponding glaciological information are compiled in Extended Data Table 1. Some of the most rapidly eroding contemporary glacial systems worldwide^{10,11} are found at the northern end of our transect. Basin-wide erosion rates range from 12 mm yr^{−1} at San Rafael Glacier in the north to 0.01 mm yr^{−1} at Funk Glacier in the polar south. As seen in Fig. 2a and b, the erosion rates show a significant correlation with latitude ($r^2 = 0.75$, $n = 13$) and mean annual temperature ($r^2 = 0.81$, $n = 13$).

Along our transect, erosion rates also increase nonlinearly with both the sliding speed and the ice flux through the ELA (Fig. 2c, d), suggesting a weak power-law relationship with both ice discharge and basal speed ($r^2 = 0.21$ and $r^2 = 0.39$, respectively). This nonlinear

relationship is in accord with theories of glacial erosion, where an increase in basal ice velocities is expected to increase both quarrying rates and the flux of debris available to abrade the bed^{17,18}. The two outliers in this general trend are Tyndall Glacier and Fourcade Glacier; excluding these two systems, the correlation between erosion and sliding improves substantially ($r^2 = 0.62$; Extended Data Fig. 1).

We caution that several sources of uncertainty in our measurements of both erosion rate and ice motion may confound simple relationships between the two parameters. For instance, the sediment yield and inferred erosion rate for Europa Glacier are abnormally low for its size, speed and climate, which probably reflects both reduced erosion and the trapping of sediments in a deep subglacial basin. Europa is the only Patagonian glacier in our data set that has not undergone substantial thinning and terminus retreat in the past 50 years³⁰, and hence we are not yet seeing the increase in sediment yield that has been observed to accompany glacial retreat^{10,11}. Moreover, at Europa Glacier, surface speeds decrease sharply with distance from the terminus and remain low 8–15 km upglacier of the terminus (see Extended Data Figs 2 and 3). Mass (ice) conservation suggests that this decrease in surface speed, combined with low surface slopes, reflects an abrupt increase in ice thickness and an extensive subglacial overdeepening. Shallow slopes and overdeepenings with steep outlets favour the storage of sediment at the bed, both reducing sediment delivery to the ice front and protecting the bed from further erosion. Moreover, there is no obvious submarine sill or moraine in the outer fjord to trap all the products of glacial erosion. Topographic controls on subglacial and proglacial sediment storage and evacuation are sources of uncertainty in all of our glacier–fjord systems, but are most pronounced for this catchment; see Methods for further discussion of the uncertainties with these estimates.

Notwithstanding the complexities inherent in our observational data, the modern erosion rates for the western Antarctic Peninsula (0.01 mm yr^{−1} to <0.1 mm yr^{−1}) are two orders of magnitude lower than the rates for the Patagonian glaciers (1 mm yr^{−1} to >10 mm yr^{−1}). This difference exists despite overlapping ice fluxes and sliding speeds. Within the western Antarctic Peninsula, our erosion rates are within the range reported previously in polar fjords in the Arctic^{12,13,23}. The erosion rates in this region also tend to increase with increasing glacier size and speed. The over 100-fold lower erosion rates in the Antarctic Peninsula suggest both less vigorous glacial erosion and less delivery of sediment generated at the bed to the ice front and

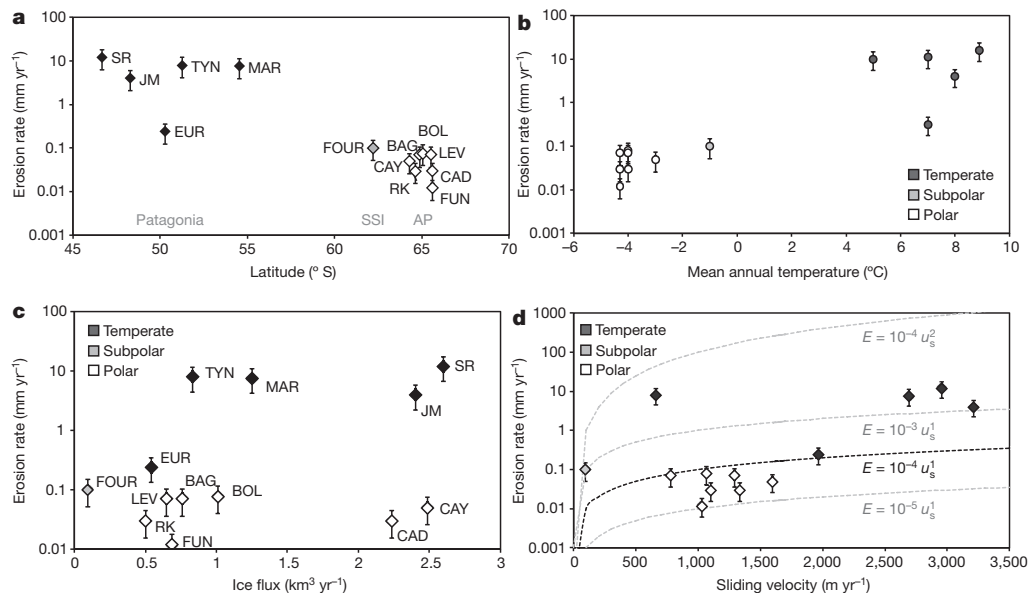


Figure 2 | Erosion rate as a function of latitude, climate and dynamics of 13 outlet glaciers. **a–d**, Erosion rate versus latitude (**a**); SSI, South Shetland Islands; AP, Antarctic Peninsula), mean annual air temperature, estimated from ref. 32 (**b**), ice flux (**c**) and basal sliding speed, with dashed lines representing commonly used forms of the ‘erosion rule’ (as labelled, the black

dashed line corresponds to the most commonly used form) in landscape-evolution models^{5–8,15,18} (**d**). Error bars denote $\pm 50\%$ uncertainty in determining sediment volume and basin-averaged bedrock erosion rate. For definitions of glacier acronyms, see Extended Data Table 1.

the fjord beyond by the subglacial hydrologic system. Further work is needed to better resolve the effect of such changes in sediment evacuation rates on the sediment yields observed over these timescales, which could confound inferred erosion rates. As suggested above, this is most probably the case for Europa Glacier.

For glaciers of similar ice flux or sliding speed across the thermal spectrum, erosion rates for polar glaciers are lower by over two orders of magnitude than for temperate glaciers with similar ice discharges (Fig. 2c). We suggest that this difference for glaciers of similar size and motion is primarily related to the abundance of meltwater accessing the bed. This is in accord with the concept that frequent and rapid fluctuations in the basal water pressure, which are more likely in regions where meltwater production is seasonal, promote both subglacial quarrying rates and efficient sediment evacuation by subglacial rivers^{17,18}. Patagonia is one such region, where the relatively warm climate and heavy precipitation augment both surface and internal melting, thereby increasing the supply of water to the glacier bed, which promotes sliding, erosion, and sediment production and evacuation⁵. For example, annual surface melt rates approach 2 m and 6 m water equivalent at Marinelli and San Rafael glaciers, respectively. In contrast, in the western Antarctic Peninsula, sub-freezing temperatures for much of the year, melt rates of approximately 0.1 m water equivalent per year and ice thicknesses of >300 m suggest that little, if any, surface melt generated can infiltrate the glaciers without refreezing³¹.

From temperate to polar settings, modern erosion rates measured from systems of similar catchment size, ice flux, tectonic history and bedrock lithology slow by over two orders of magnitude (Fig. 2). A similar decrease, which we see in space from temperate to polar settings, is widely recognized in time from data representing recent to long-term (million-year) erosion rates derived from sediment traps, cosmogenic dating and low-temperature thermochronometers from the same glaciated catchment^{10,14}. Because all glaciers worldwide have experienced generally colder-than-current climatic conditions throughout the late Quaternary period, the 100-fold decrease in long-term relative to modern erosion rates, particularly for currently temperate glaciers in Patagonia^{2,14}, may reflect (in part) the temporal averaging of warm- and cold-based conditions over the lifecycle of

these glaciers. On the basis of our ‘trading space for time’ analysis, and the 100-fold difference in erosion rates captured along our latitudinal transect, we expect a substantial acceleration in erosion rates and in sediment delivery from the glaciers of the Antarctic Peninsula region as they transition to more temperate conditions in the coming century²¹.

Recent numerical modelling efforts have successfully replicated glacial landscapes when erosion rates are assumed to scale linearly with basal sliding speed^{5–8,15,16}. Our results augment these models by providing field calibration of rates of erosion and sliding, and highlight the seldom-recognized importance of the effect of surface temperature on the erosion rate. Our findings indicate that erosion rates from glaciers under a broad range of thermal regimes are highly variable. We propose that climatic variation, more than ice dynamics, controls the temporal and spatial variability in erosion rates, and that a mean annual temperature above $0–5^{\circ}\text{C}$ (implying ample supplies of subglacial meltwater) constitutes a threshold condition for fast glacial erosion. Our findings reinforce the link between erosional processes and global climate, and help explain the role of climate change in the development of topography over glacial–interglacial timescales.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 June 2014; accepted 27 July 2015.

- Egholm, D. L., Nielsen, S. B., Pedersen, V. K. & Lesemann, J.-E. Glacial effects limiting mountain height. *Nature* **460**, 884–887 (2009).
- Thomson, S. N. *et al.* Glaciation as a destructive and constructive control on mountain building. *Nature* **467**, 313–317 (2010).
- Koons, P. O. The topographic evolution of collisional mountain belts: a numerical look at the Southern Alps, New Zealand. *Am. J. Sci.* **289**, 1041–1069 (1989).
- Stroeven, A. P., Fabel, D., Hatterstrand, C. & Harbor, J. A relict landscape in the centre of the Fennoscandian glaciation: cosmogenic radionuclide evidence of tors preserved through multiple glacial cycles. *Geomorphology* **44**, 145–154 (2002).
- Egholm, D. L., Pedersen, V. K., Knudsen, M. F. & Larsen, N. K. Coupling the flow of ice, water, and sediment in a glacial landscape evolution model. *Geomorphology* **141–142**, 47–66 (2012).
- Herman, F., Beaud, F., Champagnac, J.-D., Lemieux, J.-M. & Sternai, P. Glacial hydrology and erosion patterns: a mechanism for carving glacial valleys. *Earth Planet. Sci. Lett.* **310**, 498–508 (2011).
- Yanites, B. & Ehlers, T. A. Global climate and tectonic controls on the denudation of glaciated mountains. *Earth Planet. Sci. Lett.* **325–326**, 63–75 (2012).

8. Tomkin, J. H. Numerically simulating alpine landscapes: the geomorphic consequences of incorporating glacial erosion in surface process models. *Geomorphology* **103**, 180–188 (2009).
9. Hallet, B., Hunter, L. & Bogen, J. Rates of erosion and sediment evacuation by glaciers: a review of field data and their implications. *Global Planet. Change* **12**, 213–235 (1996).
10. Koppes, M. & Montgomery, D. The relative efficacy of fluvial and glacial erosion over modern to orogenic timescales. *Nature Geosci.* **2**, 644–647 (2009).
11. Koppes, M., Hallet, B. & Anderson, J. Synchronous acceleration of ice loss and glacier erosion, Marinelli Glacier, Tierra del Fuego. *J. Glaciol.* **55**, 207–220 (2009).
12. Szczuciński, W., Zajączkowski, M. & Scholten, J. Sediment accumulation rates in subpolar fjords—impact of post-Little Ice Age glaciers retreat, Billefjorden, Svalbard. *Estuar. Coast. Shelf Sci.* **85**, 345–356 (2009).
13. Cowton, T., Nienow, P., Bartholomew, I., Sole, A. & Mair, D. Rapid erosion beneath the Greenland Ice Sheet. *Geology* **40**, 343–346 (2012).
14. Fernandez, R. A., Anderson, J. B., Wellner, J. S. & Hallet, B. Timescale dependence of glacial erosion rates: a case study of Marinelli Glacier, Cordillera Darwin, southern Patagonia. *J. Geophys. Res.* **116**, F01020 (2011).
15. MacGregor, K. R., Anderson, R. S. & Waddington, E. D. Numerical modeling of glacial erosion and headwall processes in alpine valleys. *Geomorphology* **103**, 189–204 (2009).
16. Kessler, M. A., Anderson, R. S. & Briner, J. P. Fjord insertion into continental margins driven by topographic steering of ice. *Nature Geosci.* **1**, 365–369 (2008).
17. Hallet, B. Glacial quarrying: a simple theoretical model. *Ann. Glaciol.* **22**, 1–8 (1996).
18. Iverson, N. R. A theory of glacial quarrying for landscape evolution models. *Geology* **40**, 679–682 (2012).
19. Cowan, E. A. *et al.* Fjords as temporary sediment traps: history of glacial erosion and deposition in Muir Inlet, Glacier Bay National Park, southeastern Alaska. *Geol. Soc. Am. Bull.* **122**, 1067–1080 (2010).
20. Boldt, K. V. *et al.* Modern rates of glacial sediment accumulation along a 15° S–N transect in fjords from the Antarctic Peninsula to southern Chile. *J. Geophys. Res.* **118**, 2072–2088 (2013).
21. Cuffey, K. M., Conway, H., Hallet, B., Gades, A. & Raymond, C. F. Interfacial water in polar glaciers and glacier sliding at -17°C . *Geophys. Res. Lett.* **26**, 751–754 (1999).
22. Hooke, R. & Elverhøi, A. Sediment flux from a fjord during glacial periods, Isfjorden, Spitsbergen. *Global Planet. Change* **12**, 237–249 (1996).
23. Humphrey, N. F. & Raymond, C. F. Hydrology, erosion and sediment production in a surging glacier: Variegated Glacier, Alaska, 1982–83. *J. Glaciol.* **40**, 539–552 (1994).
24. Griffith, T. W. & Anderson, J. B. Climatic control of sedimentation in bays and fjords of the northern Antarctic Peninsula. *Mar. Geol.* **85**, 181–204 (1989).
25. DaSilva, J. L., Anderson, J. B. & Stravers, J. Seismic facies changes along a nearly continuous 24° latitudinal transect: the fjords of Chile and the northern Antarctic Peninsula. *Mar. Geol.* **143**, 103–123 (1997).
26. Hebbeln, D., Lamy, F., Mohtadi, M. & Echler, H. Tracing the impact of glacial-interglacial climate variability on erosion of the southern Andes. *Geology* **35**, 131–134 (2007).
27. Domack, E. W. & McClennen, C. E. Accumulation of glacial marine sediments in fjords of the Antarctic Peninsula and their use as late Holocene paleoenvironmental indicators. *Antarct. Res. Ser.* **70**, 135–154 (1996).
28. Rignot, E., Rivera, A. & Casassa, G. Contribution of the Patagonia Icefields of South America to sea level rise. *Science* **302**, 434–437 (2003).
29. Rau, F. & Braun, M. The regional distribution of the dry-snow zone on the Antarctic Peninsula north of 70° S. *Ann. Glaciol.* **34**, 95–100 (2002).
30. Sukakibara, D. & Sugiyama, S. Ice-front variations and speed changes of calving glaciers in the Southern Patagonia Icefield from 1984–2011. *J. Geophys. Res.* **119**, 2541–2554 (2014).
31. Lenaerts, J. T. M., van den Broeke, M. R., van den Berg, W. J., van Meijgaard, E. & Kuipers Munneke, P. A new, high-resolution surface mass balance map of Antarctica (1979–2010) based on regional atmospheric climate modeling. *Geophys. Res. Lett.* **39**, L04501 (2012).
32. Morris, E. M. & Vaughan, D. G. in *Antarctic Peninsula Climate Variability* (eds Domack, E. *et al.*) Vol. 79 of *Antarctic Research Series* 61–68. (American Geophysical Union, 2003).

Acknowledgements This research was funded by the US National Science Foundation (OPP 0338371). We thank the crews of the ice breaker RV *Nathaniel B. Palmer* and the MV *Petrel IV*, members of Waters of Patagonia, support staff from Raytheon Polar Services, and collaborators from the Centro de Estudios Científicos in Valdivia, Chile, the University of Washington, Rice University and the University of Houston for assisting in deployments, sampling and analysis of the sediment cores, bathymetric data, ice front geometries and acoustic reflection profiles collected during the cruises. We particularly thank J. Anderson, A. Rivera, M. Jaffrey, J. Evans and T. Verzone for help and logistical support in the field; R. Sylwester for his contribution to the collection of acoustic reflection profiles in Chile; C. Nittrouer, B. Forrest, C. Landowski, J. Berquist and T. Drexler for processing and analysing the sediment cores; T. Pratt for processing of acoustic profiles in Jorge Montt; J. Anderson and R. Fernandez for supporting data and discussions; C. Brookfield for editing and insight; R. Jaña at INACH for provision of Landsat imagery of the Antarctic Peninsula; and M. Jaffrey, J. Newton and A. Winter-Billington for help with statistical analyses.

Author Contributions M.K., B.H. and J.S.W., together with J. Anderson, designed the study. M.K. conducted all analyses of glaciological and erosion-rate data, and prepared the manuscript. E.R. and J.M. contributed the ice-velocity measurements. K.B. contributed the accumulation-rate results, and provided new data for Jorge Montt Glacier. J.S.W. and K.B. analysed the bathymetric data, acoustic profiles and sediment cores in the Antarctic Peninsula fjords. All authors contributed to discussions and interpretations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.K. (koppes@geog.ubc.ca).

METHODS

Geologic setting. Southern Chile and the Antarctic Peninsula share the tectonic and geologic history of the Andean orogen, the product of a long history of ocean–continent collision and ridge subduction since the mid-Palaeozoic era³³. From the Late Palaeozoic to the Tertiary period, the southern Andes and the northern Antarctic Peninsula have been in a state of tectonic compression and experiencing coeval island-arc marginal basin evolution, giving rise to a thick sequence of fore-arc sedimentary basin deposits intruded by calc-alkaline plutons³⁴. Since the Neogene period, subduction of the Nazca–Antarctic and Phoenix–Antarctic spreading ridges has produced the Patagonian and Antarctic slab windows, giving rise to the interruption of normal calc-alkaline arc volcanism, the eruption of plateau basalts in southern South America and the Antarctic Peninsula³⁴, and the broad uplift of both continents through a dynamic topographic response³⁵. These slab windows persist today, producing persistent mantle upflow, broad regional uplift and basaltic magmatism that overlie the plutons and sedimentary basin deposits³⁴. Thus, relatively resistant metasediments, intrusive granite batholiths, and mafic metavolcanics are the predominant rock types that underlie the icefields from Patagonia to the Antarctic Peninsula, with bed resistances within a relatively narrow range.

Climatic setting. The climate along our transect varies from the warm and wet sector of Patagonia in southern Chile, to a transitional climate in the South Shetland Islands, to the relatively cold and dry western Antarctic Peninsula. The study area spans almost 20° of latitude and has mean annual air temperatures (MAT) that vary by 14 °C, encompassing temperate, subpolar and polar glaciers that range from the lowest-latitude tidewater glacier in the world, San Rafael Glacier (46.6° S, MAT = 9 °C), to Cadman Glacier, Beascochea Bay (65.6° S, MAT = −5 °C). Precipitation rates at sea level range from about 7 m yr^{−1} in southern Patagonia to about 1 m yr^{−1} in the western Antarctic Peninsula.

At the northern end of our transect, westerly winds create a large E–W precipitation gradient along Chile’s southern coast. Near the core of the westerlies at 50° S, the latitude of the South Patagonian Icefield (Fig. 1b), at sea level, the mean annual temperature is approximately 7 °C and precipitation reaches 7 m yr^{−1}. To the north, at the latitude of San Rafael Glacier (46° S), precipitation decreases to about 4 m yr^{−1} and mean annual temperatures at sea level average approximately 9 °C. To the south, in the Cordillera Darwin Icefield (54° S), precipitation rates are closer to 1 m yr^{−1} and mean annual temperatures are 5 °C at sea level. Owing to the relatively warm and wet setting of Chilean Patagonia, all of the ice masses are temperate and have presumably been temperate since the early Holocene³⁶.

The South Shetland Islands lie close to the northern tip of the Antarctic Peninsula (Fig. 1a, c). Land- and marine-terminating ice masses, with a mean thickness of approximately 250 m, cover the islands. The climatic conditions vary between subpolar and temperate, with winter temperatures between −3 °C and −5 °C, and late summer temperatures above freezing³². This area has experienced 3.7 ± 2.1 °C of warming in the last century—the second-fastest warming documented worldwide³⁷. Of the total annual precipitation of about 1.2 m, approximately 0.2 m falls as rain during the summer: it contributes to substantial ablation and adds surface water to the lower reaches of the icefields, where late summer snowlines vary around approximately 150 m above sea level³⁸.

Along the western margin of the Antarctic Peninsula, the climate varies from polar to subpolar and mean temperatures range from about 0 °C in the austral summer to −8 °C to −11 °C in the winter³⁸. Average precipitation at sea level along the western Antarctic Peninsula varies from approximately 0.8 m water equivalent per year at Andvord Bay to approximately 1 m water equivalent per year at Beascochea Bay³⁸, all as snow. The narrow N–S spine of the Antarctic Peninsula, much like in the Patagonian Andes, creates a strong orographic gradient that supports a narrow plateau of ice, with steep, narrow outlet glaciers cascading to the west from the summit plateau and terminating at sea level.

The Antarctic Peninsula is one of the most rapidly warming regions in the world, with an increase in MAT of 3.7 ± 1.6 °C over the past century³². Over the last 50 years, the most rapid warming recorded in Antarctica of 5.7 ± 2.0 °C per century was measured at Faraday Station (65.2° S, 64.3° W), near Beascochea Bay³⁷. The glaciers of the western Antarctic Peninsula are already showing substantial variability in both contemporary ice fluxes and in retreat rates in response to the rapid regional warming^{39–41}. The sediment output (and hence erosion rate) is expected to increase as these glaciers accelerate and as a more robust subglacial hydrologic system flushes out sediment that may have been stored under the glaciers. Hence, we expect further changes in ice dynamics and associated increases in erosion rates as the region continues to warm, particularly as many of the outlet glaciers of the Antarctic Peninsula have already started to accelerate as climatic conditions shift to a more temperate regime³⁷.

Calculating contemporary ice fluxes and sliding speeds. The contemporary ice flux through each glacier system was reconstructed, with the exception of the glaciers mentioned below, by multiplying the cross-sectional area of the ice front

with the depth-averaged velocity near the terminus, based on surface-velocity measurements using synthetic-aperture-radar interferometry (InSAR)^{42–44} (Extended Data Fig. 2)

For the five Patagonian glaciers, we use previously published estimates of the surface speed and/or ice flux at the ELA. For the San Rafael and Marinelli glaciers, the average ice flux at the ELA over the past 50 years was calculated using a mass-balance budget model described in refs 11 and 45. The ice flux for Europa Glacier was reported in ref. 42, for Jorge Montt Glacier in ref. 46 and for Tyndall Glacier in ref. 47; all studies followed a similar approach to that described below.

For Fourcade Glacier of the South Shetland Islands, the cross-sectional area at the ELA (approximately 250 m above sea level) was measured using ice-penetrating radar (Extended Data Fig. 4). Velocity stakes were also installed at the ELA and tracked using a differential global positioning system (dGPS) over a two-week period in April 2007 to complement surface velocities with InSAR measurements. The use of short-term velocities measured at the end of the summer melt season produces greater uncertainties and may underestimate annual surface velocities, and hence ice flux, for this glacier system; that said, the maximum surface velocities of approximately 100 m yr^{−1} that we measured are within the range of surface velocities (60–150 m yr^{−1}) obtained from InSAR measurements averaged over 2007–2011. The potential for underestimation of the ice velocity, coupled with the predominantly volcanic bed lithology of King George Island, which suggests less resistance to mechanical and chemical erosion, may help explain (in part) why the erosion rate for Fourcade Glacier is high relative to ice motion (see Fig. 2 and Extended Data Fig. 1).

For the southernmost study glaciers of the western Antarctic Peninsula, where the ELAs are essentially at sea level, swath bathymetry was used from the ice breaker RV *Nathaniel B. Palmer* in April 2007 to determine the fjord width and depths along the ice fronts. The heights of the ice cliffs above waterline were also measured along each terminus using photogrammetry and navigational radar aboard the ice breaker RV *Nathaniel B. Palmer*. Together, the bathymetry and ice-cliff estimates were used to determine the cross-sectional area of the calving front, which represents the width-averaged ice thickness at the ELA (Extended Data Fig. 5).

To calculate the ice sliding velocity across the ELA of each glacier we followed the methods of ref. 44. The surface velocity measured from InSAR (or the velocity stakes) is the sum of the sliding and deformational velocities: $u_{\text{surface}} = u_s + u_d$.

Where some contribution of the flow is accommodated by deformation, the ice velocity decreases with depth at a rate that increases with the shear stress. We assume Glen’s flow law for the strain rate: $\dot{\epsilon} = A\tau^n$, where τ is the shear stress; with a temperature-dependent stiffness constant of $A = 1.7 \times 10^{-24} \text{ s}^{-1} \text{ Pa}^{-3}$ for subpolar ice at −2 °C for the glaciers of the Antarctic Peninsula and $A = 2.4 \times 10^{-24} \text{ s}^{-1} \text{ Pa}^{-3}$ for the temperate ice of the Patagonian glaciers; and an exponent $n = 3$.

The sliding velocity u_s is then the difference between the surface velocity u_{surface} and the depth-integrated strain rate:

$$u_s = u_{\text{surface}} - \int_0^H A(\rho g \sin(\theta))^n h^n dh$$

where

$$A(\rho g \sin(\theta))^n = A\tau^n = \frac{du}{dh} = \dot{\epsilon}$$

is the strain rate, ρ is the density of ice, θ is the surface slope and H is the width-averaged ice thickness at the ELA. The deformation component assumes deformation by simple shear, and does not include the effect of longitudinal stress gradients or drag along valley walls. Using the known ice thickness H and the surface velocities that we measured, the maximum internal deformational velocities can be tightly constrained, and contribute no more than about 20–30 m yr^{−1} (for ice thicknesses of 400–800 m) to the total motion. Given the high surface speeds (greater than 1,000 m yr^{−1} in most cases) and modest thicknesses (a few hundred metres or less) of all of these glaciers, the ice surface velocity predominantly reflects basal sliding (93%–100% of surface speed, see Extended Data Table 1) and very little internal deformation for all glaciers in our study.

The ice discharge (flux) Q at the ELA is therefore $Q_{\text{ELA}} = Fu_s W_{\text{ELA}} H_{\text{ELA}}$, where W_{ELA} and H_{ELA} are the width and ice thickness at the ELA, respectively, and $F = (W_{\text{ELA}} + H_{\text{ELA}})/W_{\text{ELA}}$ is a semi-elliptical shape factor.

Calculating the contemporary basin-averaged bedrock erosion rate. Two cruises aboard the ice breaker RV *Nathaniel B. Palmer* during 2005 and 2007, as well as three separate cruises aboard small vessels in 2004 (Tyndall), 2006 (San Rafael) and 2010 (Jorge Montt) provided the sediment cores, acoustic reflection profiles and bathymetric data from which the sediment yields were quantified and erosion rates estimated. Cruise NBP0505 mapped the Europa and Marinelli fjords in southern Chile. Cruise NBP0703 mapped Maxwell Bay in the South Shetland

Islands, as well as six fjords along the western Antarctic Peninsula from 62° S to 65° S. Within each fjord, 3.5-kHz CHIRP and/or 100-Hz bubble pulser acoustic reflection profiles and multibeam swath bathymetry were collected (see ref. 20 for the bathymetric maps). On the basis of the detailed bathymetry and sub-bottom profiles, the sediment depocentres closest to, or abutting, the glacier termini were identified, and kasten cores were collected therein (see Extended Data Fig. 2), from which sediment accumulation rates were quantified using ^{210}Pb chronology^{20,27}.

Owing to substantial differences in accumulation rates between the fjords, we used two different methods to quantify the sediment yields from the glaciers. For temperate systems, where sediment accumulation rates close to the ice fronts of several metres per year have been observed^{11,19,20,25} and the glaciers have retreated >5 km over the last century from known terminus positions^{11,14,19,47,48,49}, the sediment yields over the past century were reconstructed from the total sediment volume mapped in the acoustic reflection profiles in the proximal fjord, within the extent of the innermost moraine, dated to >1950 AD from aerial photos^{11,48,49}. In these areas of rapid retreat and rapid sedimentation, with the exception of Europa Glacier and the distal basin of Marinelli Glacier (beyond the 1960 moraine), the short (1–2 m) kasten cores did not capture any measurable accumulation rate signal (they showed uniform, low excess ^{210}Pb in the upper 1–2 m of sediment). Hence, for these basins with rapid accumulation and a known retreat history, seismic mapping of the total sediment volume in the inner fjord provides the best measure of the modern sediment yield. When estimating yields for these glaciers, we follow the methods of refs 11, 14, 19, 48 and 49. As in similar studies, we assume that all the semi-transparent, laminated and hummocky sediment visible above a strong reflector in the sub-bottom profiles represent the post-retreat sedimentary package, deposited as the glacier terminus retreated across the fjord basin mostly in the second half of the twentieth century. Triangulated irregular network interpolation of the acoustic reflection profiles, which was used to produce gridded sediment thicknesses across the inner basins, introduces at most an 18% error in total sediment thickness, with the error increasing with both distance between profile tracks and spatial variability in the sediment thicknesses. Including a user error of 1%–2% in picking sediment depths from the seismic profiles and another 5% error in applying a median seismic velocity of $1,700 \text{ m s}^{-1}$ for glaciomarine muds, we estimate that the total error in determining the sediment volume in the proximal basins from the acoustic reflection profiles is $\pm 25\%$.

In contrast, at Europa Glacier and the glaciers of the western Antarctic Peninsula, the temporal resolution of retreat and accumulation were too low to be captured in the acoustic reflection profiles. The sediment yields for Europa Glacier, the distal basin of Marinelli Glacier and the glaciers of the Antarctic Peninsula were therefore reconstructed from sediment deposition rate, on the basis of ^{210}Pb chronology in the upper 1 m of the cores²⁰; the rate was assumed to be uniform over the areal extent of the ice-proximal depocentres mapped in the swath bathymetry. ^{210}Pb chronology has been used successfully in the ice-distal regions of temperate fjords^{19,20,49,50} and throughout polar fjords^{12,20,27}; it provides the sediment accumulation rate in each depocentre averaged over the past century (about five ^{210}Pb half-lives). In a few fjords (Cierva Cove and Collins Bay), the profiles show little or no excess ^{210}Pb , suggesting either no substantial accumulation (which goes against all circumstantial evidence, including a thick drape of soft, unconsolidated sediments) or (which is more likely) a rapid accumulation of >1 m of sediment that did not sequester appreciable ^{210}Pb from the fjord water column²⁰. For these glaciers, it was not possible to derive erosion rates. From the acoustic reflection profiles (archived in ref. 51), it was clear that in most instances: (1) the sediment layers were of uniform thickness across each depocentre, which leads us to assume that once the sediment has been diffusively redistributed along the bottom of the fjord, accumulation is spatially uniform across each depocentre, despite the large variation in deposition rates observed in temperate fjords^{19,20,49,50}; and (2) the bedrock highs surrounding these depocentres were devoid of a sediment drape, and hence we assume that all the sediment delivered by the glacier collected in the basin lows, whether by direct rain-out of fine sediment from the water column or through gravitational redistribution of sediment that had collected on the steep flanks of the fjord and sills. We recognize that in all cases these depocentres are not entirely closed systems, and that there is potential for some of the sediment to by-pass the most proximal depocentre through entrainment in the water column, or to 'leak' from the depocentre through resuspension and remobilization downfjord. If a substantial portion of the sediment is transported beyond the fjord, our estimates of sediment yield within the proximal depocentres would not capture the total volume of sediment being eroded and delivered by the glacier. To address this, in four of the fjords (Marion Cove (Maxwell Bay), Andvord Bay, Flandres Bay and Beascochea Bay) sediment cores were also collected in the middle/outer fjord depocentres, from which modern accumulation rates were also estimated using ^{210}Pb and ^{137}Cs chronologies²⁷. From these outer basins, we

estimate that 46%–54% of the total sediment flux from the glaciers are by-passing the proximal depocentre. Hence, the sediment yields we are measuring from the proximal basins represent approximately half of the total flux; we assumed this to be true for all fjords that do not have prominent outer sills, we doubled the estimated sediment yields that are based on only the proximal basins.

The varying distribution of accumulation within each fjord and the percentage of sediment lost beyond the proximal basins, present a distinct challenge for estimating basin-averaged erosion rates in 'open' fjord systems. In addition, we recognize the limitations of using one or two point measurements of sediment accumulation rates from each basin to calculate total sediment yields. The sediment cores in the most proximal basin were generally collected more than 500 m from the ice front (with the exception of Trooz Glacier, see ref. 20), and probably do not capture the highest accumulation rates that are expected close to the ice front. Thus, our assumption that the measured sediment accumulation rates^{20,27} are representative of the entire proximal basin depocentre probably also underestimates the total volume of sediment delivered by the glacier, and thus the erosion rate. We also assume that the contribution of non-terrestrial, biogenic material to the fjord sediment is small, less than 10% of the overall sediment volume deposited^{20,27}.

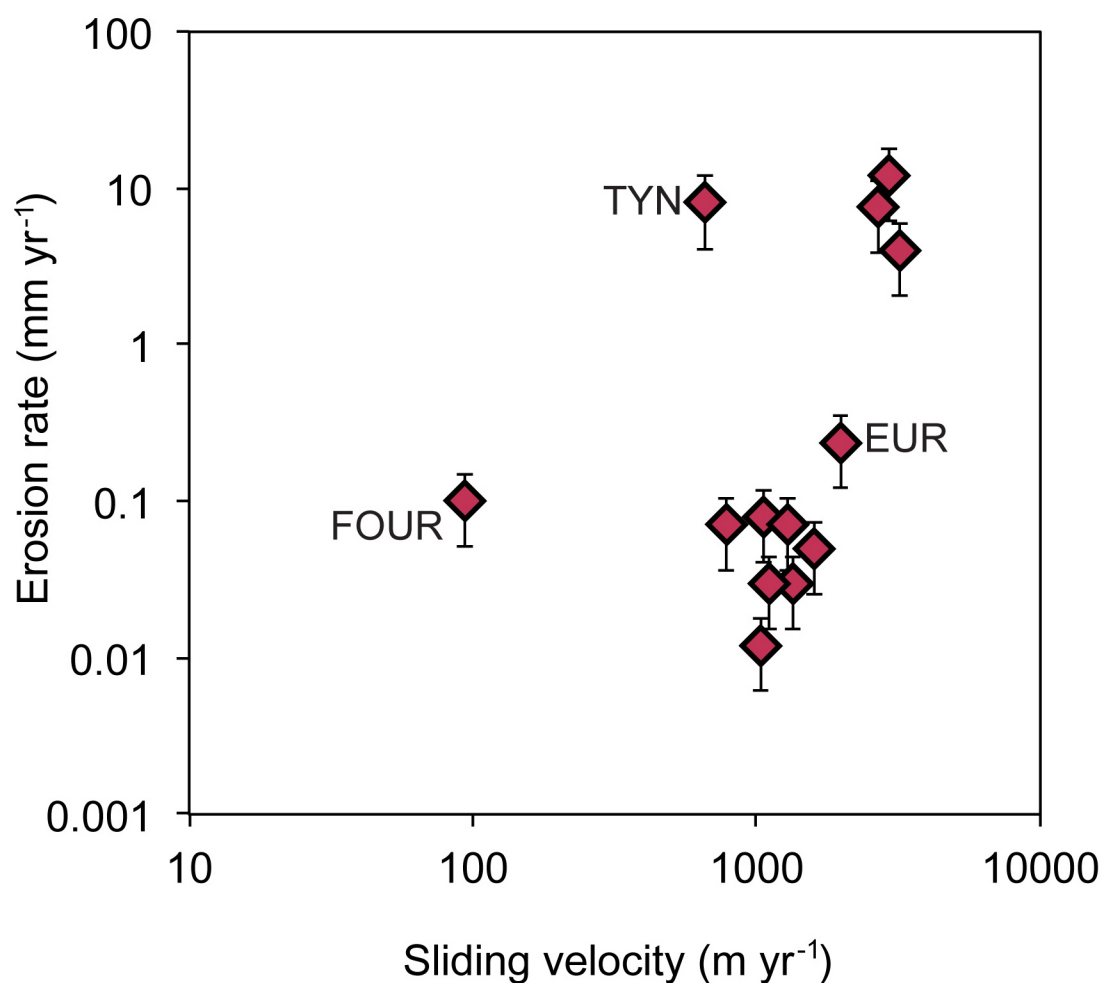
At Marinelli Glacier, the accumulation rates in the distal basin were small enough that we were able to compare the sediment yields calculated using the total sediment volume from the proximal basin with those using ^{210}Pb chronology from a core in the distal depocentre (12 km from the current terminus, and outside of a moraine that the glacier retreated from around 1960^{11,14,20}); see Extended Data Fig. 6. For this glacier system, the ^{210}Pb measurement of distal accumulation produced sediment yields that were about a third (36%) of the yield that was estimated from the total sediment volume in the proximal basin (see Extended Data Table 1). However, in this instance, comparison of the two methods is complicated by real differences in local and basin-wide measurements and in proximity to the ice front, and hence proximity to the highest rates of accumulation, over the centennial timescale of measurement. Nevertheless, the comparison of the two methods to measure sediment yields at this one location suggests that any methodological biases produce a factor of 2–3 difference in estimates at most, but clearly cannot account for the differences of over two orders of magnitude in the sediment yields, and by inference erosion rates, between temperate and polar systems.

Using both measurement approaches, the 100-year-averaged sediment yield (in cubic metres per year) is then converted to a basin-wide sediment production rate per unit area by dividing the yield by the glacier catchment area, measured from 2005 Landsat imagery, the 2000 Shuttle Radar Topography Mission digital elevation model (SRTM DEM; for Patagonia) and the British Antarctic Survey's Radarsat 200-m digital elevation model (for the Antarctic Peninsula); see Extended Data Fig. 2. To convert the sediment production rate to a bedrock erosion rate, the former is multiplied by the ratio between the dry bulk density of glaciomarine sediment ($1,300 \text{ kg m}^{-3}$, the median density measured from the sediment cores) and that of crystalline metasedimentary and igneous bedrock (approximately $2,700 \text{ kg m}^{-3}$). Using this approach, we derive a centennially averaged, basin-averaged bedrock erosion rate for each glacier catchment. The range of bedrock and sediment densities introduces an additional uncertainty in the calculation of the bedrock erosion rate of up to 12%. Hence, combined with uncertainties in our assumption of the terrestrial origin of all sediment, not accounting for spatial variations in the sediment accumulation rate and in the areal extent of deposition, the cumulative known uncertainty in our calculated basin-wide erosion rates approaches 38% for the temperate fjords and 50% for Europa and the polar fjords, and all are probably underestimates of the total erosion and sediment produced by the glaciers.

The relationship between erosion and sliding. To compare our findings regarding the relationship between erosion and sliding (the 'erosion rule') to what has been used in numerical models (Fig. 2d), we employed a nonlinear least-squares regression analysis using the 'nls' package in R⁵² to estimate the two constants of proportionality, K_g and n , from the erosion rule $E = K_g u_s^n$, where u_s and E are our observations of basal sliding speed and estimates of erosion rate for each glacier, respectively. We computed the mean and standard deviation of the erosion rates and ice velocities. We then ran the nonlinear least-squares model within the 95% confidence interval of the observed data, assuming both the power-law distribution, and an exponential distribution of the form $E = \alpha e^{bu_s}$. We also used a linear least-squares fit for comparison. The best-fit model (residual standard error, $\text{RSE} = 0.00335$; residual sum-of-squares error, $\text{RSS} = 0.0001246$; $r^2 = 0.39$) using the nonlinear least-squares method for all glaciers in the data set ($n = 13$) returned a power-law distribution with coefficients $K_g = 5.2 \times 10^{-8}$ and $n = 2.34$. Using the same suite of nonlinear models and excluding the two outlier data—Tyndall Glacier and Fourcade Glacier (see

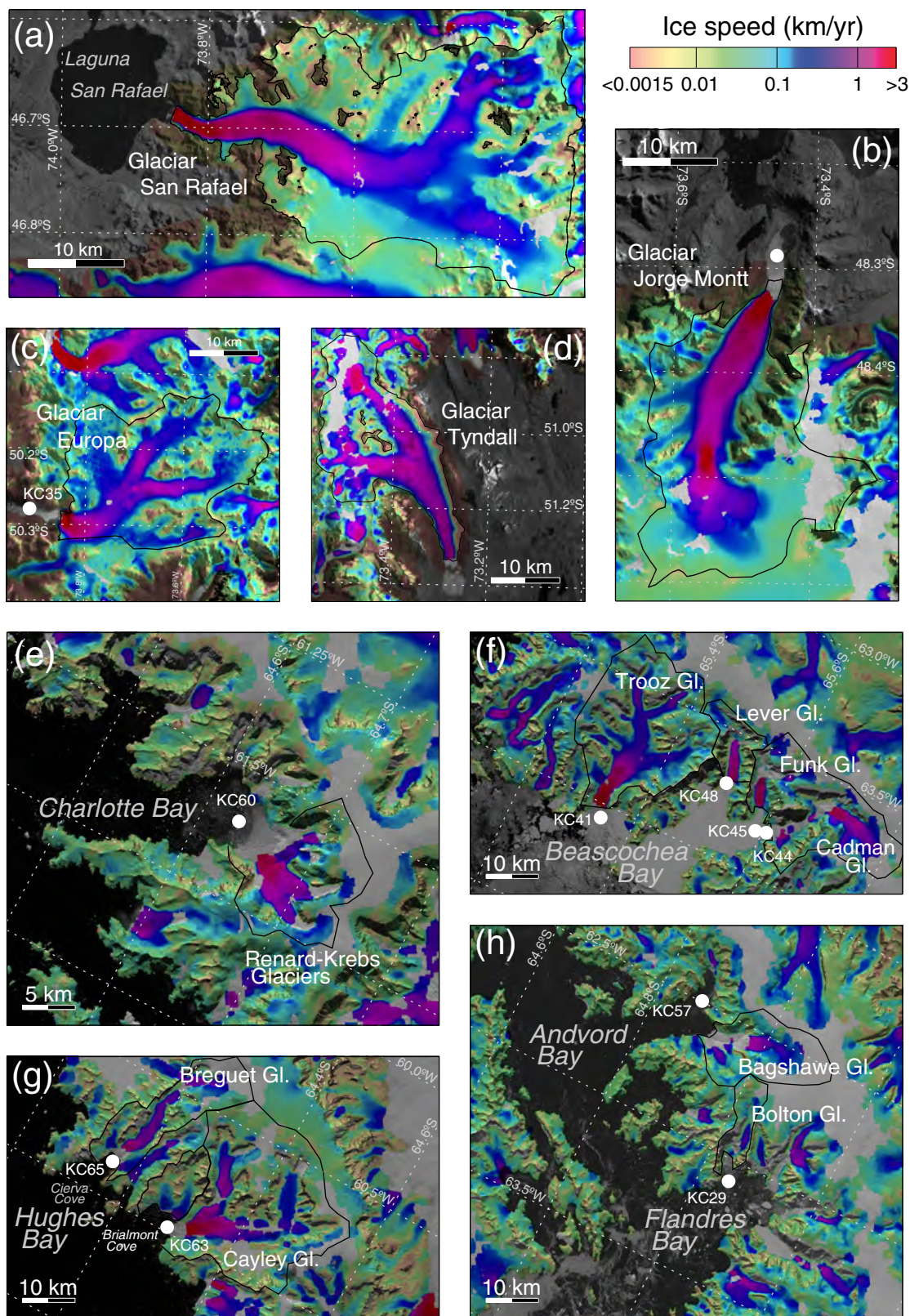
Extended Data Fig. 1)—also resulted in a best-fit power-law distribution ($r^2 = 0.62$) with coefficients $K_g = 5.3 \times 10^{-9}$ and $n = 2.62$.

33. Dalziel, I. W. D. *et al.* The Scotia arc: genesis, evolution, global significance. *Annu. Rev. Earth Planet. Sci.* **41**, 767–793 (2013).
34. Breitsprecher, K. & Thorkelson, D. J. Neogene kinematic history of Nazca–Antarctic–Phoenix slab windows beneath Patagonia and the Antarctic Peninsula. *Tectonophysics* **464**, 10–20 (2009).
35. Guillaume, B. *et al.* Dynamic topography control on Patagonian relief evolution as inferred from low temperature thermochronology. *Earth Planet. Sci. Lett.* **364**, 157–167 (2013).
36. Mayr, C. *et al.* Holocene variability of the Southern Hemisphere westerlies in Argentinean Patagonia (52° S). *Quat. Sci. Rev.* **26**, 579–584 (2007).
37. Vaughan, D. G. *et al.* Recent rapid regional climate warming on the Antarctic Peninsula. *Clim. Change* **60**, 243–274 (2003).
38. Turner, J. *et al.* Antarctic climate change during the last 50 years. *Int. J. Climatol.* **25**, 279–294 (2005).
39. Cook, A. J., Fox, A. J., Vaughn, D. G. & Ferrigno, J. G. Retreating glacier fronts on the Antarctic Peninsula over the past half-century. *Science* **308**, 541–544 (2005).
40. Cook, A. J., Vaughn, D. G., Luckman, A. & Murray, T. A new Antarctic Peninsula glacier basin inventory and observed area changes since the 1940s. *Antarct. Sci.* **26**, 614–624 (2014).
41. Pritchard, H. D. & Vaughn, D. G. Widespread acceleration of tidewater glaciers on the Antarctic Peninsula. *J. Geophys. Res.* **112**, F03S29 (2007).
42. Forster, R. R., Rignot, E., Isacks, B. L. & Jezek, K. C. Interferometric observations of Glaciares Europa and Penguin, Hielo Patagonico Sur, Chile. *J. Glaciol.* **45**, 325–337 (1999).
43. Mougnot, J. & Rignot, E. Ice motion of the Patagonian Icefields of South America: 1984–2014. *Geophys. Res. Lett.* **42**, 1441–1449 (2015).
44. Rignot, E., Mougnot, J. & Scheuchl, B. Ice flow of the Antarctic Ice Sheet. *Science* **333**, 1427–1430 (2011).
45. Koppes, M., Conway, H., Rasmussen, L. A. & Chernos, M. Deriving mass balance and calving variations from reanalysis data and sparse observations, Glaciar San Rafael, northern Patagonia, 1950–2005. *Cryosphere* **5**, 791–808 (2011).
46. Rivera, A., Corripio, J., Bravo, C. & Cisternas, S. Glaciar Jorge Montt (Chilean Patagonia) dynamics derived from photos obtained by fixed cameras and satellite image feature tracking. *Ann. Glaciol.* **53**, 147–155 (2012).
47. Raymond, C. *et al.* Retreat of Glaciar Tyndall, Patagonia, over the last half-century. *J. Glaciol.* **51**, 239–247 (2005).
48. Koppes, M., Sylwester, R., Rivera, A. & Hallet, B. Variations in sediment yield over the advance and retreat of a calving glacier, Laguna San Rafael, North Patagonian Icefield. *Quat. Res.* **73**, 84–95 (2010).
49. Boldt, K. V. *Fjord sedimentation during the rapid retreat of tidewater glaciers: observations and modeling*. PhD thesis, Univ. Washington (2014).
50. Jaeger, J. M., Nittrover, C. A., Scott, N. D. & Milliman, J. D. Sediment accumulation along a glacially impacted mountainous coastline: north-east Gulf of Alaska. *Basin Res.* **10**, 155–173 (1998).
51. Carbotte, S. M. *et al.* *Antarctic Multibeam Bathymetry and Geophysical Data Synthesis: an on-line digital data resource for marine geoscience research in the Southern Ocean*. Open-file Report No. 2007-1047-SRP-002 (US Geological Survey, 2007).
52. Baty, F. *et al.* A toolbox for nonlinear regression in R: the package nlstools. *J. Stat. Softw.* **66**, 5 (2015).



Extended Data Figure 1 | Erosion rate versus sliding velocity for 13 outlet glaciers. A log-log plot, showing a general power-law relationship between erosion and basal ice motion, and two outliers: Fourcade Glacier (FOUR) and Tyndall Glacier (TYN). The nonlinear least-squares best-fit estimate using all

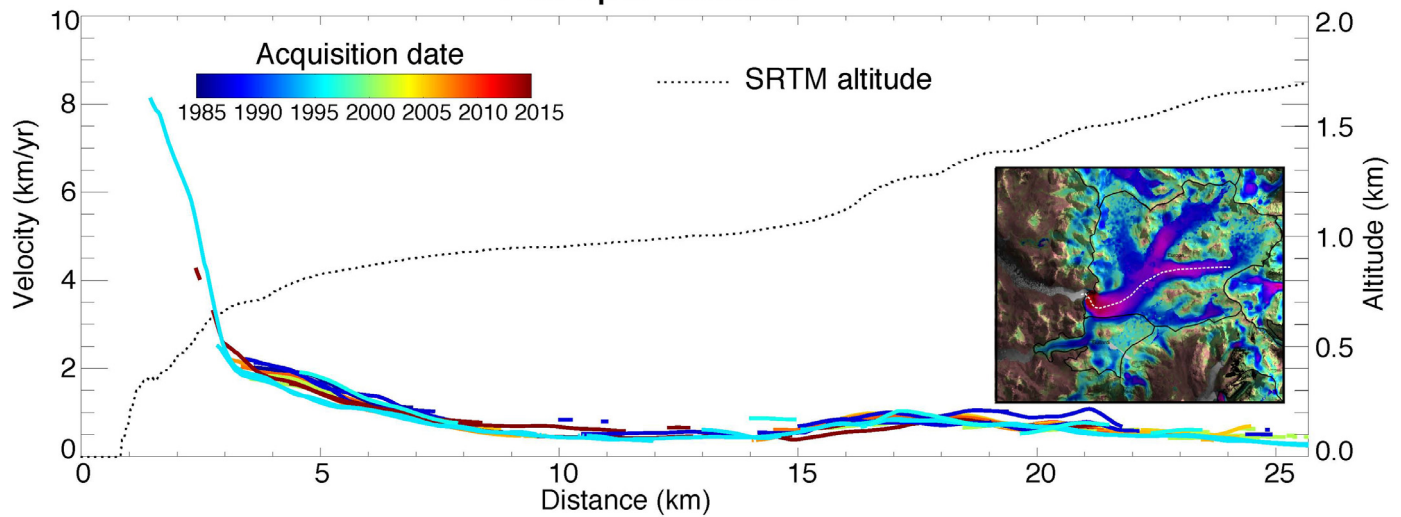
glaciers yields an exponent $n = 2.34$ and intercept $K_g = 5.2 \times 10^{-8}$ ($r^2 = 0.39$); excluding the two outliers, the fit improves, with $n = 2.62$ and intercept $K_g = 5.3 \times 10^{-9}$ ($r^2 = 0.62$).



Extended Data Figure 2 | Ice motion for outlet glaciers of Patagonia and western Antarctic Peninsula. a–h, Glacier catchment areas (within the black outline) with InSAR-derived ice velocities (in km yr^{-1}) from 2007–2008 superimposed (indicated by the colour scale). The InSAR velocity maps are modified from data from refs 43 and 44. White dots indicate the location of

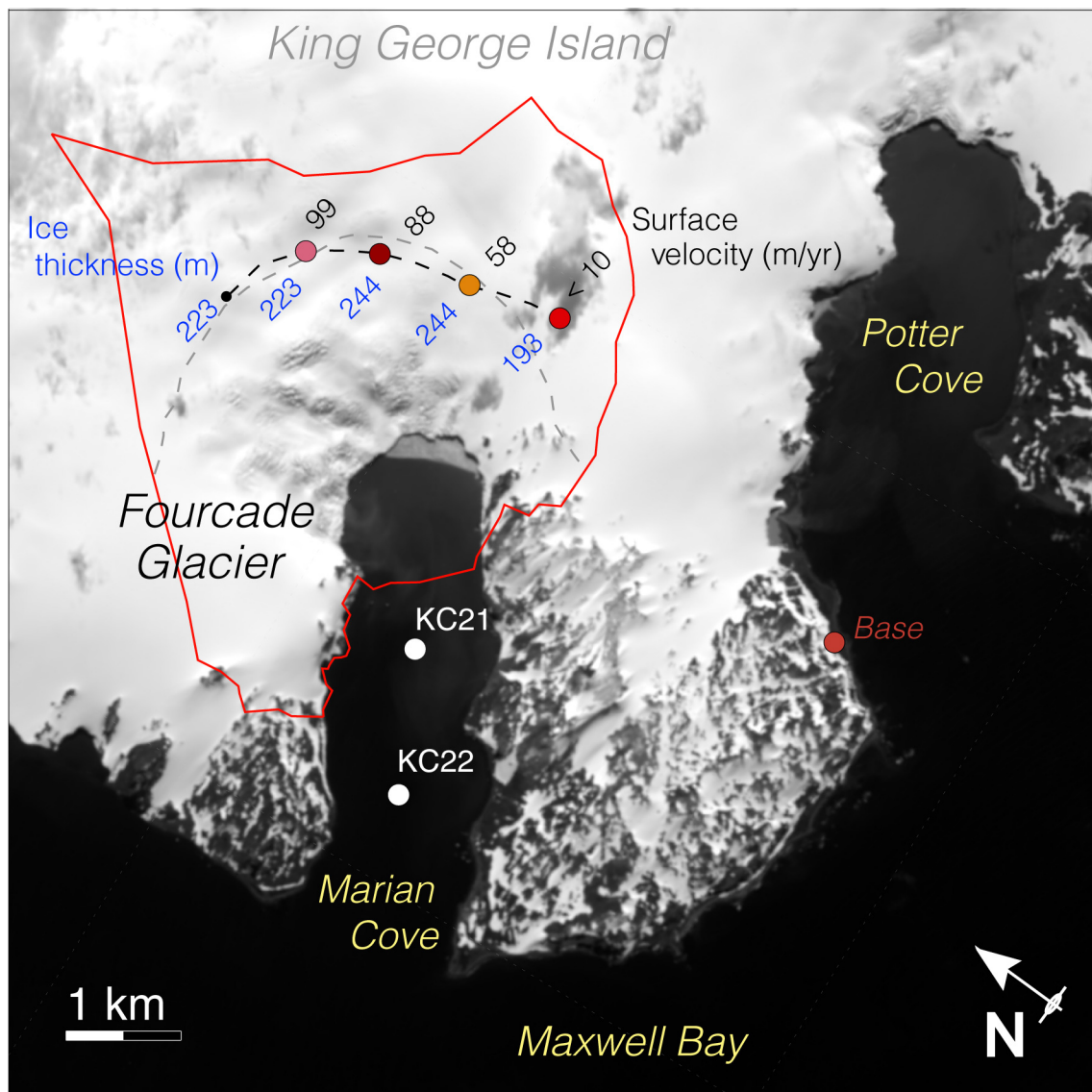
the cores in ice-proximal depocentres from which accumulation rates were measured (see refs 20, 27 and 49). Catchment areas shown are San Rafael (a), Jorge Montt (b), Europa (c), Tyndall (d), Charlotte Bay (e), Beascochea Bay (f), Hughes Bay (g), and Andvord and Flandres Bays (h).

Europa Glacier



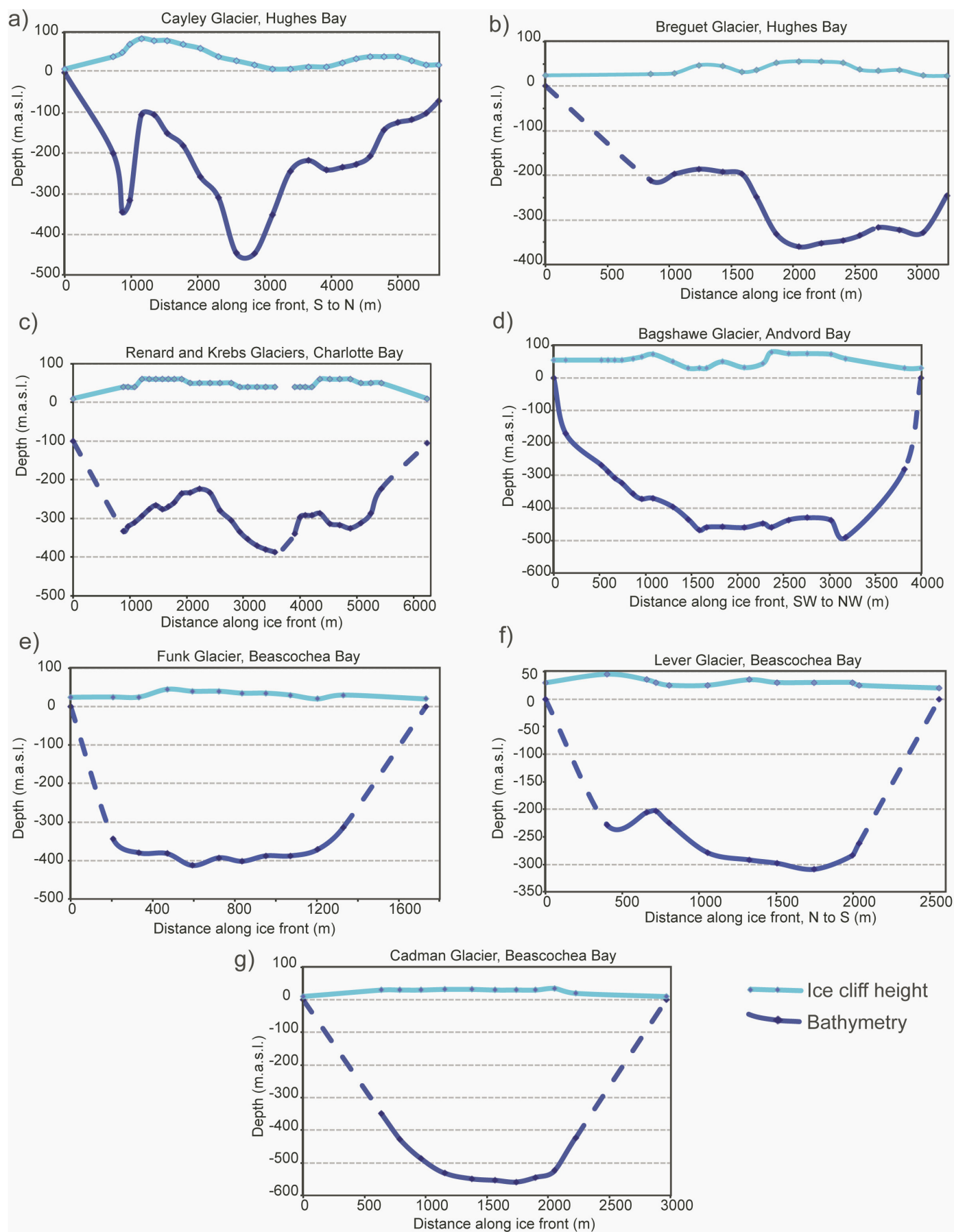
Extended Data Figure 3 | Surface elevation and time series of surface velocity along the central flowline for Europa Glacier, South Patagonian Icefield. The black dashed line is the elevation profile from the terminus, derived from the 2001 SRTM DEM. Flow speeds were measured along the

centreline from InSAR repeat image pairs (see ref. 43), coloured according to the year the data were acquired. Inset is Extended Data Fig. 2c, with the overlaid white dashed line indicating the centreline of the glacier.



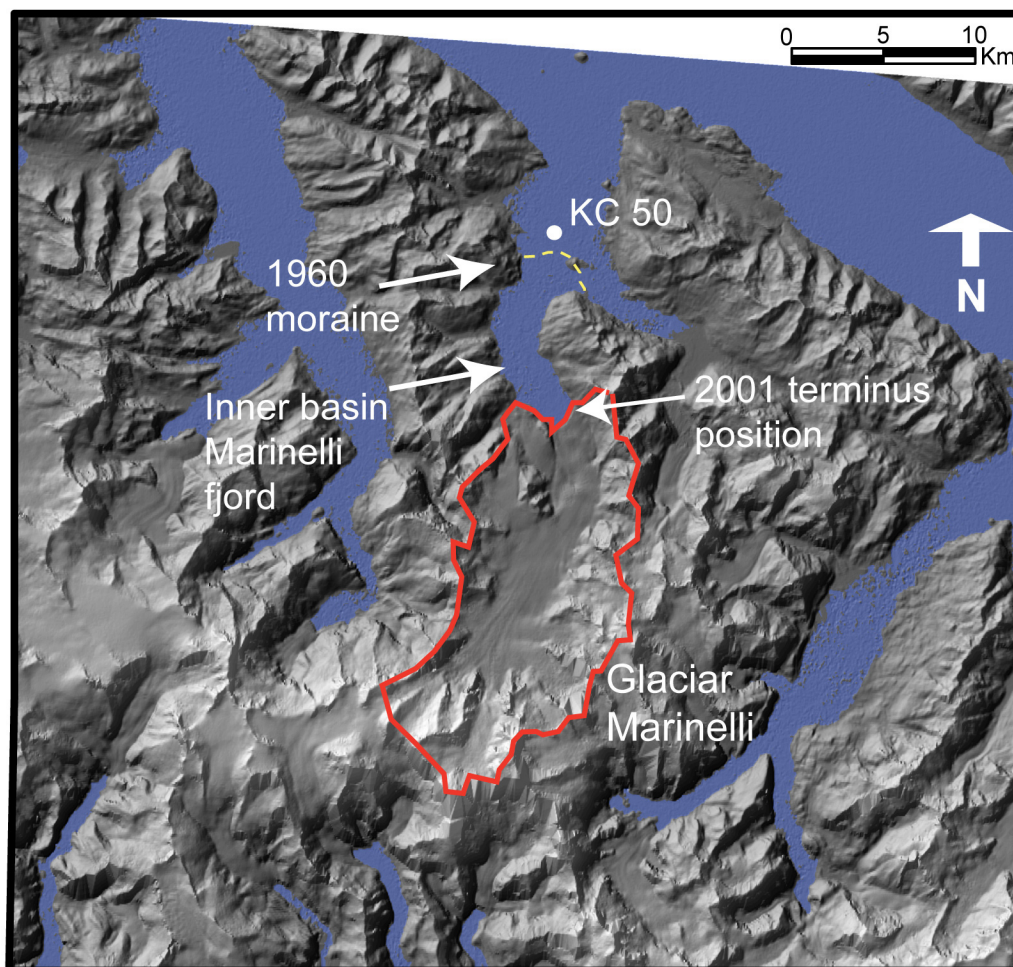
Extended Data Figure 4 | Ice thickness and surface velocity across Fourcade Glacier, King George Island. The red line indicated the glacier catchment area in 2007; the black dashed line shows the path of the ice-penetrating radar; and the grey dashed line is the ELA (approximately 250 m above sea level).

Surface velocities from dGPS of velocity stakes (April 2007) are in black and ice thickness measurements from ice-penetrating radar are in blue. Base indicates location of dGPS base station.



Extended Data Figure 5 | Ice-front cross-sectional areas of the polar glaciers of the western Antarctic Peninsula. a–g, The light blue lines are the ice cliff heights above the water line and the dark blue lines are the submarine ice faces from the swath bathymetry; m.a.s.l., metres above sea level. Dashed lines indicate interpolated ice thicknesses between known points. For all

glaciers, the ELA is located at the calving front. Measurements are for Cayley Glacier, Hughes Bay (a), Breguet Glacier, Hughes Bay (b), Renard and Krebs glaciers, Charlotte Bay (c), Bagshawe Glacier, Andvord Bay (d), Funk Glacier, Beascochea Bay (e), Lever Glacier, Beascochea Bay (f) and Cadman Glacier, Beascochea Bay (g).



Extended Data Figure 6 | Glacier and fjord catchment area for Marinelli Glacier, Cordillera Darwin Icefield. The glacier catchment area in 2009 is indicated by the red line; the location of Little Ice Age moraine from which the glacier terminus retreated around 1960 is indicated by the yellow dashed line; the inner basin of the fjord where acoustic reflection profiles

captured total sediment volume since 1960 is indicated by the appropriate arrow (see ref. 11); and the white dot indicates the location of the sediment core in the distal depocentre from which the distal accumulation rate was measured (see ref. 20).

Extended Data Table 1 | Glacier characteristics and corresponding erosion rates in Chilean Patagonia and the Antarctic Peninsula.

| GLACIER | Location | Lat. | Long. | Cross-sectional area | | Width of calving front ² | MAX surface velocity ³ | Ratio basal: surface velocity | ICE FLUX | dx/dt ⁴ | Sediment Accumulation Rate ⁵ | Proximal depocenter area ⁶ | Erosion Rate ^{7,8} |
|------------------------|----------------|-------|-------|-----------------------------|-----------------------------------|-------------------------------------|-----------------------------------|-------------------------------|---------------------|--------------------|---|---------------------------------------|-----------------------------|
| | | | | Catchment Area ¹ | Glacier frontal area ² | | | | | | | | |
| | | °S | °W | km ² | km ² | km | m/yr | % | km ³ /yr | m/yr | mm/yr | km ² | mm/yr |
| <i>Patagonia</i> | | | | | | | | | | | | | |
| San Rafael (SR) | NPI | 46.66 | 73.82 | 728 | 0.42 | 2.2 | >2800 | 100 | 0.75-2.3 | -88 | 643 | 13.9 | 12 ± 4 |
| Jorge Montt (JM) | SPI | 48.3 | 73.5 | 500 | 0.44 | 1.6 | >3600 | 100 | 2.1 | -278 | >400 | 4 | 4 ± 1.2 |
| Europa (EUR) | SPI | 50.28 | 73.92 | 409 | 0.27 | 1.1 | 1970 | 100 | 0.54 | | 11.4 | 8.98 | 0.24 ± 0.12 |
| Tyndall (TYN) | SPI | 51.25 | 73.27 | 418 | 0.44 | 1.9 | 700 | 100 | 0.83 | -123 | | 1.8 | 8 ± 2.4 |
| Marinelli (MAR) | CDI | 54.39 | 69.59 | 154 | 0.23 | 1.8 | >1000 | 100 | 0.5-1.17 | -289 | >500 | 21.7 | 7.6 ± 2.3 |
| | | | | | | | | | | | 23.6 | 42 | 2.71 ± 1.3 |
| <i>S. Shetland Is.</i> | | | | | | | | | | | | | |
| Fourcade (FOUR) | Marion Cove | 62.21 | 58.75 | 241 | 0.90 | 4.1 | 100 | 96 | 0.09 | -6.86 | 6.6, 2.8 | 0.26, 0.29 | 0.10 ± 0.05 |
| <i>Antarctica Pen.</i> | | | | | | | | | | | | | |
| Breguet | Cierva Cove | 64.16 | 60.85 | 261 | 0.74 | 3.5 | 880 | 96 | 1.64 | -4.9 | uniform | 1.1 | n/a |
| Cayley (CAY) | Brailmont Cove | 64.30 | 60.98 | 720 | 1.44 | 5.6 | 1670 | 97 | 2.48 | -1.7 | 15 | 2.47 | 0.05 ± 0.03 |
| Renard-Krebs (RK) | Charlotte Bay | 64.66 | 61.62 | 118 | 0.95 | 4.3 | 1200 | 95 | 0.5 | -4.9 | 2.8 | 1.51 | 0.03 ± 0.02 |
| Bagshawe (BAG) | Andvord Bay | 64.91 | 62.61 | 241 | 1.66 | 5.8 | 1450 | 93 | 0.76 | 7.7 | 5.6 | 2.06 | 0.07 ± 0.03 |
| Bolton (BOL) | Flandres Bay | 65.07 | 62.97 | 62 | 0.62 | 2.6 | 1100 | 98 | 1.01 | -55 | 2.8 | 3.77 | 0.08 ± 0.04 |
| Trooz | Collins Bay | 65.33 | 63.97 | 633 | 0.94 | 3.3 | >1360 | 98 | 1.57 | 37.6 | uniform | 2.26 | n/a |
| Lever (LEV) | Beascochea Bay | 65.51 | 63.69 | 177 | 1.20 | 2.6 | >800 | 98 | 0.65 | -4.2 | 6.4 | 2.23 | 0.07 ± 0.03 |
| Funk (FUN) | Beascochea Bay | 65.58 | 63.78 | 142 | 0.58 | 1.7 | 1075 | 97 | 0.69 | -4.1 | 2.0 | 2.47 | 0.01 ± 0.005 |
| Cadman (CAD) | Beascochea Bay | 65.61 | 63.82 | 306 | 1.15 | 3.0 | >1500 | 93 | 2.23 | 16.3 | 7.0 | 1.65 | 0.03 ± 0.02 |

CDI, Cordillera Darwin Icefield; NPI, North Patagonian Icefield; SPI, South Patagonian Icefield.

¹ Glacier basin area was measured from 2005 Landsat 7, 2013 Landsat Enhanced Thematic Mapper imagery and the British Antarctic Survey's Radarsat 200-m DEM (see Extended Data Fig. 2).

² Width of calving-front and ice-front cross-sectional area was measured from swath bathymetry and ice cliff height (see Extended Data Fig. 5).

³ Maximum and width-averaged surface ice flow speed at ELA was measured from InSAR velocity map (see Extended Data Fig. 2).

⁴ Retreat rates (dx/dt) for individual glaciers in the latter half of the twentieth century were reported in ref. 39.

⁵ Sediment accumulation rates were calculated from ²¹⁰Pb decay profiles in top 1 m of cores, reported in refs 20 and 27. Accumulation rates, and therefore erosion rates, from vertically uniform ²¹⁰Pb profiles could not be determined.

⁶ Depocentre basin area was measured from multibeam swath bathymetry and 3.5-kHz sub-bottom acoustic reflection profiles (see ref. 20 for bathymetric maps and data repository in ref. 51 for acoustic profiles).

⁷ Bedrock erosion rate was calculated by dividing the centennial sediment yield (the product of sediment accumulation rate and depocentre area) by the glacier catchment area, taking into account the dry bulk densities of glaciomarine sediment (on average, 1.3 g cm⁻³) and crystalline bedrock (on average, 2.7 g cm⁻³). Erosion rates for San Rafael, Jorge Montt, Tyndall and Marinelli glaciers (inner basin) were calculated using the total sediment volume deposited since 1960 in the basin closest to the ice front, measured from acoustic reflection profiles and repeat bathymetry (see refs 11, 48, and 49).

⁸ Estimates of centennial sediment yield for the Antarctic Peninsula and Europa glaciers were doubled from the yields measured in the proximal depocentre to account for downfjord losses in accumulation, as estimated from accumulation rates in the middle and outer basins of Marion Cove (Maxwell Bay), Andvord Bay (ref. 27), Flandres Bay and Beascochea Bay.

Declining global warming effects on the phenology of spring leaf unfolding

Yongshuo H. Fu^{1,2}, Hongfang Zhao¹, Shilong Piao^{1,3,4}, Marc Peaucelle⁵, Shushi Peng^{1,5}, Guiyun Zhou⁶, Philippe Ciais^{1,5}, Mengtian Huang¹, Annette Menzel^{7,8}, Josep Peñuelas^{9,10}, Yang Song¹¹, Yann Vitisse^{12,13,14}, Zhenzhong Zeng¹ & Ivan A. Janssens²

Earlier spring leaf unfolding is a frequently observed response of plants to climate warming^{1–4}. Many deciduous tree species require chilling for dormancy release, and warming-related reductions in chilling may counteract the advance of leaf unfolding in response to warming^{5,6}. Empirical evidence for this, however, is limited to saplings or twigs in climate-controlled chambers^{7,8}. Using long-term *in situ* observations of leaf unfolding for seven dominant European tree species at 1,245 sites, here we show that the apparent response of leaf unfolding to climate warming (S_T , expressed in days advance of leaf unfolding per °C warming) has significantly decreased from 1980 to 2013 in all monitored tree species. Averaged across all species and sites, S_T decreased by 40% from 4.0 ± 1.8 days °C⁻¹ during 1980–1994 to 2.3 ± 1.6 days °C⁻¹ during 1999–2013. The declining S_T was also simulated by chilling-based phenology models, albeit with a weaker decline (24–30%) than observed *in situ*. The reduction in S_T is likely to be partly attributable to reduced chilling. Nonetheless, other mechanisms may also have a role, such as ‘photo-period limitation’ mechanisms that may become ultimately limiting when leaf unfolding dates occur too early in the season. Our results provide empirical evidence for a declining S_T , but also suggest that the predicted strong winter warming in the future may further reduce S_T and therefore result in a slowdown in the advance of tree spring phenology.

The phenology of spring leaf unfolding influences regional and hemispheric-scale carbon balances², the long-term distribution of tree species⁹, and plant–animal interactions¹⁰. Changes in the phenology of spring leaf unfolding can also exert biophysical feedbacks on climate by modifying the surface albedo and energy budget^{11,12}. Recent studies have reported substantial advances in spring phenology as a result of warming in most Northern Hemisphere regions^{1,3,4}. Climate warming is projected to further increase¹³, but the future evolution of the phenology of spring leaf unfolding remains uncertain—in view of the imperfect understanding of how the underlying mechanisms respond to environmental stimuli^{12,14}. In addition, the relative contributions of each environmental stimulus, which together define the apparent temperature sensitivity of the phenology of spring leaf unfolding (S_T), may also change over time^{6,8,15}. An improved characterization of the variation in phenological responses to spring temperature is thus valuable, provided that it addresses temporal and spatial scales relevant for regional projections.

Many studies have reported advanced spring leaf unfolding which matches warming trends over recent decades^{1,3,4}. However, there is still debate regarding the linearity of the leaf unfolding response to climate warming^{6,7}. Recent experimental studies of warming using saplings have shown that S_T weakens as warming increases⁷. Experimental

manipulation of temperature for saplings or twigs, however, might elicit phenological responses that do not accurately reflect the response of mature trees^{16,17}. We therefore investigated the temporal changes in S_T in adult trees monitored *in situ* and exposed to real-world changes in temperature and other climate variables. These long-term data series were obtained across Central Europe from the Pan European Phenology Project (<http://www.pep725.eu/>). Data were collected from 1,245 sites for seven dominant tree species (see Methods and the distribution of the sites in Extended Data Fig. 1). The aims of our analysis are to determine the temporal changes in S_T at the species level during 1980–2013, a period during which Europe has substantially warmed¹³, and to relate these changes in S_T to differences in other physiological and environmental factors.

For each species at each observation site, we first determined the pre-season length as the period before leaf unfolding for which the partial correlation coefficient between leaf unfolding and air temperature was highest (see Methods). We used a gridded climate data set, including daily maximum and minimum air temperature, precipitation and absorbed downward solar radiation, with a spatial resolution of 0.25° (approximately 25 km)¹⁸. The optimal length of the pre-season ranged between 15 days and 4 months across the seven species (Extended Data Fig. 2), in agreement with earlier results^{11,14}. We then calculated the average temperature during the pre-season for each year at each site and calculated S_T using ordinary least squares linear regression for the entire period and for two 15-year periods, 1980–1994 and 1999–2013, that had a slight difference in pre-season lengths (Extended Data Fig. 3a). The leaf unfolding dates were negatively correlated with the pre-season temperature, with a mean linear correlation coefficient of -0.61 ± 0.16 , determined using the pre-season defined from the time period 1980–2013. Almost all individual tree-level correlations were negative (99.7%) and the vast majority of these correlations was statistically significant at $P < 0.05$ (93.4%) (Extended Data Fig. 4). Consistent with previous studies^{1,4}, the timing of leaf unfolding substantially advanced in all species for 1980–2013, with an average advancing rate of 3.4 ± 1.2 days °C⁻¹ across all species sites (hereafter, a positive value indicates advancement) (Fig. 1a). But the surprising result is that S_T significantly decreased by 40.0% from 4.0 ± 1.8 days °C⁻¹ during 1980–1994 to 2.3 ± 1.6 days °C⁻¹ during 1999–2013 ($t = -37.3$, $df = 5,473$, $P < 0.001$) (Fig. 1b). All species show similar significant decreases in S_T (Fig. 1a), although the extent of reduction was species-specific. For example, *Aesculus hippocastanum* (see caption to Fig. 1 for English common names) had the largest decrease in S_T (-2.0 days °C⁻¹), whereas S_T decreased only slightly (but still significantly) in *Fagus sylvatica* (-0.9 days °C⁻¹) (Fig. 1a). Similar results were also obtained using a fixed pre-season length deter-

¹Sino-French Institute for Earth System Science, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China. ²Centre of Excellence PLECO (Plant and Vegetation Ecology), Department of Biology, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium. ³Key Laboratory of Alpine Ecology and Biodiversity, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100085, China. ⁴Center for Excellence in Tibetan Earth Science, Chinese Academy of Sciences, Beijing 100085, China. ⁵Laboratoire des Sciences du Climat et de l'Environnement, CEA CNRS UVSQ, Gif-sur-Yvette 91190, France. ⁶School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China. ⁷Ecoclimatology, Technische Universität München, Freising 85354, Germany. ⁸Technische Universität München, Institute for Advanced Study, Lichtenbergstraße 2a, 85748 Garching, Germany. ⁹CREAF, Cerdanyola del Vallès, Barcelona 08193, Catalonia, Spain. ¹⁰CSIC, Global Ecology Unit CREA-FCI-UB, Cerdanyola del Vallès, Barcelona 08193, Catalonia, Spain. ¹¹Department of Atmospheric Sciences, University of Illinois, Urbana, Illinois 61801, USA. ¹²University of Neuchâtel, Institute of Geography, Neuchâtel 2000, Switzerland. ¹³WSL Swiss Federal Institute for Forest, Snow and Landscape Research, Neuchâtel 2000, Switzerland. ¹⁴WSL Institute for Snow and Avalanche Research SLF, Group Mountain Ecosystems, Davos 7260, Switzerland.

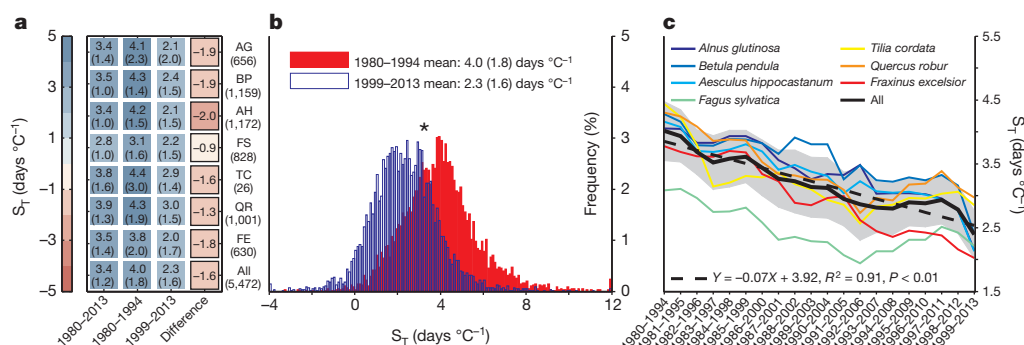


Figure 1 | Changes of apparent temperature sensitivity of leaf unfolding (S_T) over time. **a**, Species-specific S_T and s.d. (in brackets) across all sites in three periods and its difference between 1999–2013 and 1980–1994. The S_T was determined using the pre-season fixed at the time period 1980–2013 and using ordinary least squares linear regression. The colour scale indicates the magnitude of S_T . AG, alder (*Alnus glutinosa*); BP, silver birch (*Betula pendula*); AH, horse chestnut (*Aesculus hippocastanum*); FS, beech (*Fagus sylvatica*); TC, lime (*Tilia cordata*); QR, oak (*Quercus robur*); FE, ash (*Fraxinus excelsior*).

mined either in the time period 1980–1994 or in 1999–2013 (Extended Data Fig. 3b, c). The declining S_T could, however, also have been an artefact resulting from the ‘encroachment’ of leaf unfolding dates into the pre-season period that was used to calculate the temperature sensitivity. We therefore calculated the number of ‘encroachment days’ and found it is very small compared to the pre-season length even in the warmest period (Extended Data Fig. 3d, e). Because the timescale of the analysis could affect the estimates of S_T ¹⁹, we also calculated S_T using 10-year intervals (instead of 15-year intervals) and found consistent results, S_T significantly decreased between the 1980s and the last decade for all species except *Tilia cordata* (Extended Data Fig. 5a). We further calculated S_T with a 15-year moving window from 1980 to 2013 and found a significant decrease ($P < 0.01$) for each of the seven species (Fig. 1c). S_T decreased by an average of 0.7 days °C^{−1} per decade across all species. Similar results were also reached when a 10-year interval was used (Extended Data Fig. 6). These results suggest a remarkable reduction in the response of leaf unfolding to the ongoing climate warming in all studied tree species in Central Europe.

As there is no single accepted theory to account for the decreased S_T over the period 1980–2013, we propose three mutually non-exclusive hypotheses: (1) adaptation to increased variance in spring temperature, (2) photoperiodic limitations (due to earlier leaf unfolding) overriding temperature controls, or (3) reduced duration and/or sum of cold temperatures during dormancy, a ‘lost chilling’ mechanism.

The first hypothesis relates to possible effects of an increased variance in temperature. A recent study reported substantial spatial differences in S_T , with smaller absolute values at sites with a higher variance of local spring temperature²⁰. Trees may indeed develop conservative strategies (or higher phenological plasticity) of spring leaf unfolding in places where temperature fluctuates more, for example, in order to avoid spring frost damage²¹. The observed declining S_T could therefore partly result from an increase in the variance in spring temperature. However, the variance in spring temperature only significantly increased at sites of two species and decreased for all the other species except *Fraxinus excelsior* (Fig. 2a). This suggests that increased variance in spring temperature cannot account for the decreased S_T . We further studied the fluctuations in daily mean temperature and diurnal temperature amplitude ($T_{\max} - T_{\min}$) over the pre-season for the two periods 1980–1994 and 1999–2013, and for three groups of sites with comparable mean annual temperature (MAT). The fluctuations in daily temperature and diurnal temperature amplitude during the pre-season were similar during the two time periods between which S_T declined (Extended Data Fig. 7), suggesting that altered temperature variability is not an obvious cause for the declining apparent temperature sensitivity of leaf unfolding.

The number of sites for each species are in brackets below the species name. **b**, The distribution of S_T across all species and sites in two different periods and the mean S_T and s.d. (in brackets). The asterisk indicates a significant difference of S_T between the two periods at $P < 0.05$. **c**, Temporal change of S_T for individual species and for combined totals for all species across all sites with a 15-year moving window from 1980 to 2013. The black line indicates the average across all species, and the grey area indicates one s.d. either side of the mean. The dotted line indicates the linear regression.

Precocious leaf unfolding in warm springs may increase the risk of late frost events for trees²¹. To overcome this risk during warm springs, many species have evolved a protective mechanism related to photoperiod²², which hinders the warming response when days are still short and the risk for subsequent frost events is thus high. Our second, alternative, hypothesis to account for the observed decrease in S_T in recent decades is therefore a change in the relationship between chilling accumulation and heat requirement, due to the shortening days as warming advances leaf unfolding. However, we did not observe changes in S_T with latitude, neither across all species, nor for individual species (Extended Data Fig. 8), as one may expect if photoperiod was a strong co-limitation of leaf unfolding. Nonetheless, we have no evidence to exclude photoperiod as a controlling mechanism for the decline of S_T as different populations may have different genetic adaptations to photoperiod²³. In addition, the lack of relation between

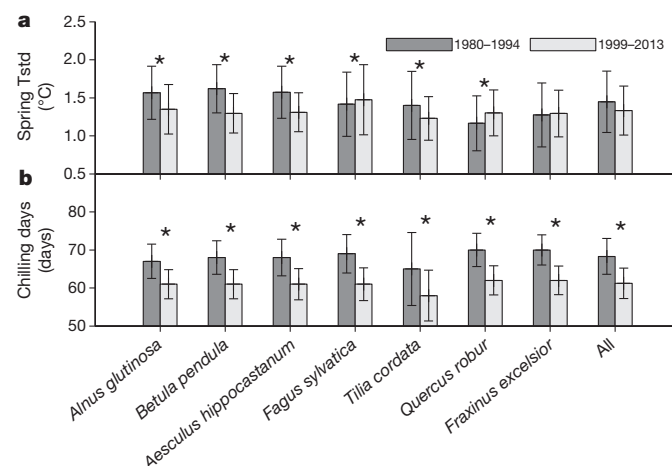


Figure 2 | Changes of chilling and spring temperature variation between 1980–1994 and 1999–2013. **a**, **b**, Species-specific Tstd (the standard deviation of mean spring temperature) (**a**) and chilling accumulation (**b**) across all sites over two periods, 1980–1994 and 1999–2013. The Tstd was calculated as the s.d. of mean spring temperature during the pre-season over these two periods. The pre-season was defined as the period before leaf unfolding for which the correlation coefficient between leaf unfolding and temperature was highest. The chilling accumulation was calculated as chilling days when daily temperature was between 0 °C and 5 °C from 1 November to the average date of leaf unfolding. The asterisks indicate significant differences at $P < 0.05$.

S_T and latitude may have been because the response of spring phenology to photoperiod can be associated with many confounding factors, such as tree age¹⁷, successional niche²³ (although there is some contradictory evidence⁸), xylem anatomy²⁴ or chilling conditions⁸. We can therefore not conclude that photoperiod did not influence S_T , but how it might directly or indirectly affect spring phenology still remains unclear and is currently under debate^{6,15,22}.

The third hypothesis to explain the decreased S_T is based on the control of spring phenology by cold temperatures during the dormancy period. In general, temperate and boreal trees require a certain amount of heat (heat requirement) after they come out of the rest period to initiate leaf unfolding in spring²⁵. This heat requirement is met sooner during warmer springs, which explains the advance of leaf unfolding in response to global warming. The heat requirement, however, is negatively correlated with chilling^{7,8,25}, that is, the accumulation of cold temperatures during the dormancy period. As the dormancy period warmed during the study period, the accumulated chilling is progressively reduced, thereby increasing the heat requirement and slowing down the advance of leaf unfolding. The net effect of lost chilling can thus be a reduced S_T . This effect may be further exacerbated by the nonlinearity of the negative correlation between the heat requirement and the accumulation of chilling^{7,25}.

To test this hypothesis, we calculated the accumulation of chilling that was defined as the sum of days when daily air temperature was within the range between 0 °C and 5 °C from 1 November in the year before leaf unfolding (see Methods), and found a significant decrease ($P < 0.001$) in chilling accumulation for all species (Fig. 2b). Chilling accumulation was 10% lower for 1999–2013 than for 1980–1994. Chilling accumulation was also significantly decreased with a 15-year moving window (Extended Data Fig. 9a) and when defined by different temperature thresholds (Extended Data Fig. 9b). To further test the importance of the ‘lost chilling’ hypothesis, we applied three chilling-based phenology models to the data (see Supplementary Information). All three models captured the declining S_T after their calibration at each site and their integration with observed climate forcing, irrespective of species (Fig. 3). The modelled relative reductions in S_T between the two periods 1980–1994 and 1999–2013 were, however, smaller than the observed decline, that is, simulated S_T was reduced by 28.8%, 24.4% and 30.4% for the sequential, parallel and unified chilling-based models, respectively, whereas the observed S_T was reduced by 40.0%. This may suggest that either other protective or adaptive mechanisms, such as photoperiod or adaptation mechanisms, are affecting the decline in S_T or that the three models do not completely accurately represent all chilling mechanisms. There are also uncertainties related to the 0.25° gridded climate product that may not represent local air temperature at each site (snow effects, shading, slope, elevation). Furthermore, using the unified model, we applied idealized step-wise increases of winter temperature over the period 1980–2013 by +1 °C to +5 °C, and consistently obtained a decrease in S_T induced by the loss of chilling in these idealized tests (Extended Data Fig. 9c). However, we did not find marked differences in S_T between years with more chilling days and years with less chilling days (Extended Data Fig. 10a–c), which can probably be explained by the different climate conditions between years with more and less chilling days. For example, the relatively high spring radiation sum in years with less chilling days might buffer the effects of less chilling days (Extended Data Fig. 10d), and eventually result in a similar S_T , but this remains a matter of speculation. Clearly, further studies are needed to support these inferences and their role in the control over phenology. Overall, these results support the third hypothesis that the decline in chilling accumulation is, at least partly, driving the decline in S_T , although the possible constraint of photoperiod/radiation could not be excluded.

Changes in spring phenology associated with climate warming have direct effects on regional and global carbon cycling¹². Studies have reported that an extension of the growing season can increase the

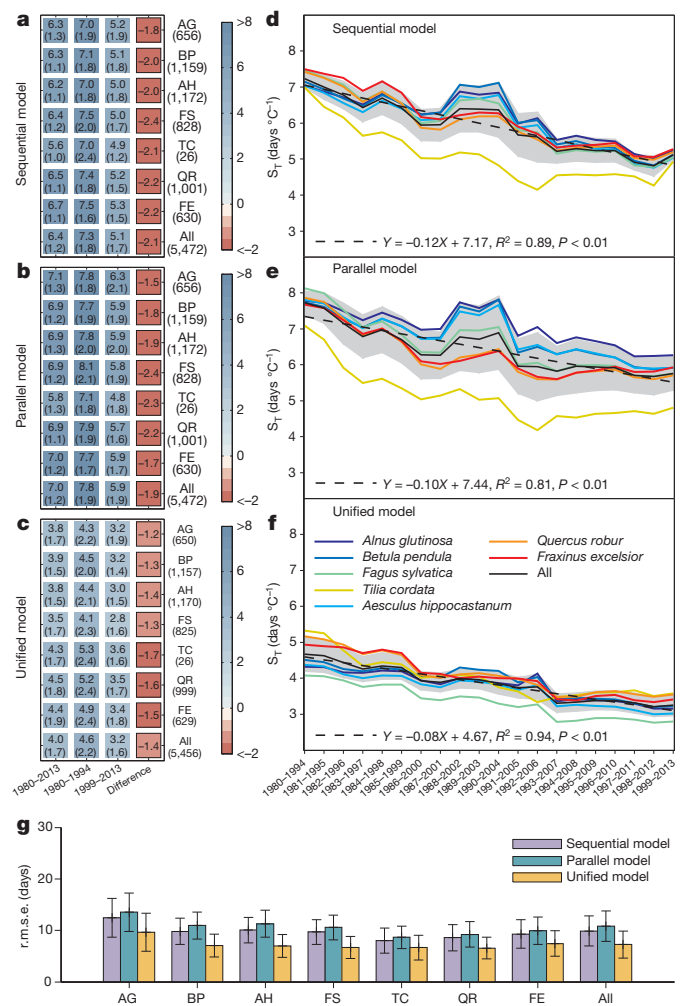


Figure 3 | Changes of modelled apparent temperature sensitivity of leaf unfolding. **a–c,** As in Fig. 1a, the panels show the modelled species-specific S_T , including the s.d. (in brackets), across all sites during three periods and its difference between 1999–2013 and 1980–1994 for the sequential model (**a**), parallel model (**b**) and unified model (**c**). **d–f,** As in Fig. 1c, the panels show the modelled temporal change of S_T for individual species and for combined totals for all species across all sites with a 15-year moving window from 1980 to 2013 for the sequential model (**d**), parallel model (**e**) and unified model (**f**). **g,** The model performance. The S_T was determined using the pre-season fixed at the time period 1980–2013 and using ordinary least squares linear regression. The colour scale indicates the magnitude of S_T . r.m.s.e., root mean square error; AG, alder (*Alnus glutinosa*); BP, silver birch (*Betula pendula*); AH, horse chestnut (*Aesculus hippocastanum*); FS, beech (*Fagus sylvatica*); TC, lime (*Tilia cordata*); QR, oak (*Quercus robur*); FE, ash (*Fraxinus excelsior*). The number of sites for each species are in brackets under the species name.

photosynthetic production of forests by 0.5–1% per day^{26–28}. We found that the apparent sensitivity of spring phenology to warming for seven temperate tree species in Central Europe has declined significantly as winter and spring temperatures increased over the past three decades. These findings indicate that the early spring phenologically driven increases in carbon uptake may slow down for temperate forests under future conditions of climate warming. On the other hand, the declining apparent temperature sensitivity of spring phenology may be beneficial for the trees. Extreme climatic events have dramatically increased in recent decades, especially warm winters and springs²⁹, and the decreased apparent temperature sensitivity would thus reduce the risk of late spring frost damage by avoiding premature leaf unfolding.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 February; accepted 21 August 2015.

Published online 23 September 2015.

1. Menzel, A. *et al.* European phenological response to climate change matches the warming pattern. *Glob. Change Biol.* **12**, 1969–1976 (2006).
2. Myneni, R. C., Keeling, C. D., Tucker, C. J., Asrar, G. & Nemani, R. R. Increased plant growth in the northern high latitudes from 1981 to 1991. *Nature* **386**, 698–702 (1997).
3. Peñuelas, J. & Filella, I. Responses to a warming world. *Science* **294**, 793–795 (2001).
4. Fu, Y. S. H. *et al.* Recent spring phenology shifts in western Central Europe based on multiscale observations. *Glob. Ecol. Biogeogr.* **23**, 1255–1263 (2014).
5. Yu, H. Y., Luedeling, E. & Xu, J. C. Winter and spring warming result in delayed spring phenology on the Tibetan Plateau. *Proc. Natl Acad. Sci. USA* **107**, 22151–22156 (2010).
6. Chuine, I., Morin, X. & Bugmann, H. Warming, photoperiods, and tree phenology. *Science* **329**, 277–278 (2010).
7. Fu, Y. S. H., Campioli, M., Deckmyn, G. & Janssens, I. A. Sensitivity of leaf unfolding to experimental warming in three temperate tree species. *Agric. For. Meteorol.* **181**, 125–132 (2013).
8. Laube, J. *et al.* Chilling outweighs photoperiod in preventing precocious spring development. *Glob. Change Biol.* **20**, 170–182 (2014).
9. Chuine, I. Why does phenology drive species distribution? *Phil. Trans. R. Soc. B* **365**, 3149–3160 (2010).
10. Zohner, C. M. & Renner, S. S. Common garden comparison of the leaf-out phenology of woody species from different native climates, combined with herbarium records, forecasts long-term change. *Ecol. Lett.* **17**, 1016–1025 (2014).
11. Peñuelas, J., Rutishauser, T. & Filella, I. Phenology feedbacks on climate change. *Science* **324**, 887–888 (2009).
12. Richardson, A. D. *et al.* Climate change, phenology, and phenological control of vegetation feedbacks to the climate system. *Agric. For. Meteorol.* **169**, 156–173 (2013).
13. *Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Field, C. B. *et al.*) (Cambridge Univ. Press, 2014).
14. Piao, S. *et al.* Leaf onset in the northern hemisphere triggered by daytime temperature. *Nature Commun.* **6**, 6911 (2015).
15. Way, D. A. & Montgomery, R. A. Photoperiod constraints on tree phenology, performance and migration in a warming world. *Plant Cell Environ.* **38**, 1725–1736 (2015).
16. Wolkovich, E. M. *et al.* Warming experiments underpredict plant phenological responses to climate change. *Nature* **485**, 494–497 (2012).
17. Vitasse, Y. Ontogenic changes rather than difference in temperature cause understory trees to leaf out earlier. *New Phytol.* **198**, 149–155 (2013).
18. Beer, C. *et al.* Harmonized European long-term climate data for assessing the effect of changing temporal variability on land-atmosphere CO₂ fluxes. *J. Clim.* **27**, 4815–4834 (2014).
19. Badeck, F. W. *et al.* Responses of spring phenology to climate change. *New Phytol.* **162**, 295–309 (2004).
20. Wang, T. *et al.* The influence of local spring temperature variance on temperature sensitivity of spring phenology. *Glob. Change Biol.* **20**, 1473–1480 (2014).
21. Vitasse, Y., Lenz, A. & Körner, C. The interaction between freezing tolerance and phenology in temperate deciduous trees. *Front. Plant Sci.* **5**, 541 (2014).
22. Körner, C. & Basler, D. Phenology under global warming. *Science* **327**, 1461–1462 (2010).
23. Basler, D. & Körner, C. Photoperiod sensitivity of bud burst in 14 temperate forest tree species. *Agric. For. Meteorol.* **165**, 73–81 (2012).
24. Hunter, A. F. & Lechowicz, M. J. Predicting the timing of budburst in temperate trees. *J. Appl. Ecol.* **29**, 597–604 (1992).
25. Harrington, C. A., Gould, P. J. & St Clair, J. B. Modeling the effects of winter environment on dormancy release of Douglas-fir. *For. Ecol. Manage.* **259**, 798–808 (2010).
26. Kimball, J. S. *et al.* Satellite radar remote sensing of seasonal growing seasons for boreal and sub-alpine evergreen forests. *Remote Sens. Environ.* **90**, 243–258 (2004).
27. Piao, S. *et al.* Growing season extension and its impact on terrestrial carbon cycle in the Northern Hemisphere over the past 2 decades. *Glob. Biogeochem. Cycles* **21**, GB3018 (2007).
28. White, M. A., Running, S. W. & Thornton, P. E. The impact of growing-season length variability on carbon assimilation and evapotranspiration over 88 years in the eastern US deciduous forest. *Int. J. Biometeorol.* **42**, 139–145 (1999).
29. Rahmstorf, S. & Coumou, D. Increase of extreme events in a warming world. *Proc. Natl Acad. Sci. USA* **108**, 17905–17909 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements This study was supported by the National Natural Science Foundation of China (41125004 and 31321061), the 111 Project (B14001), and National Youth Top-notch Talent Support Program in China. Y.H.F. is supported by an FWO Pegasus Marie Curie Fellowship. I.A.J., P.C. and J.P. acknowledge support from the European Research Council through Synergy grant ERC-2013-SyG-610028 “IMBALANCE-P” and A.M. acknowledges support through the (FP7/2007-2013)/ERC grant 282250 “E3-Extreme Event Ecology”. I.A.J. acknowledges support from the University of Antwerp Centre of Excellence “GCE”. The authors acknowledge all members of the PEP725 project for providing the phenological data.

Author Contributions Y.H.F. and H.Z. contributed equally to this work. S.Pi., Y.H.F. and I.A.J. designed the research; H.Z., Y.H.F., M.P., S.Pe. and G.Z. performed the analysis; Y.H.F., S.Pi. and I.A.J. drafted the paper; and Y.H.F., S.Pi., I.A.J., H.Z., M.P., S.Pe., G.Z., P.C., M.H., A.M., J.P., Y.S., Y.V. and Z.Z. contributed to the interpretation of the results and to the writing of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.Pi. (slpiao@pku.edu.cn).

METHODS

Data sets. *In situ* phenological observations were obtained from the Pan European Phenology network (<http://www.pep725.eu/>), which provides an open European phenological database comprising multiple plant phenological records. We selected the records of leaf unfolding dates for seven tree species at 1,245 sites for 1980–2013 from sites in an area stretching from north Germany to the Adriatic Sea (see Extended Data Fig. 1). The leaf unfolding dates were defined according to the BBCH code (Biologische Bundesanstalt, Bundessortenamt und Chemische Industrie). Dates were excluded from the analysis when the trees flushed later than the end of June. The daily mean air temperature of each site was derived from a gridded climate data set of daily maximum and minimum temperature at 0.25° spatial resolution (approximately 25 km)¹⁸. We also applied another climate forcing data set (CRU-NCEP v5, with spatial resolution of 0.5° and temporal resolution of 6 h, (<http://dods.extra.cea.fr/data/p529viov/cruncep/>)), and returned very similar results (Extended Data Fig. 5c).

Data reporting. No statistical methods were used to predetermine sample size.

Analysis. The relevant period for leaf unfolding is several months before the phenological event¹, and periods differ among species and locations. To remove covariate effects of precipitation and radiation on leaf unfolding, we applied a partial correlation analysis to determine the optimal length of the pre-season for each species at each site³⁰. The optimal pre-season for each species at each station was defined as the period (with 15-day steps) before the mean leaf unfolding date for which the partial correlation coefficient between leaf unfolding and air temperature was highest during 1980–2013. Using a similar method, we also defined pre-season for two 15-year periods (for example, 1980–1994, 1999–2013) to further assess the robustness of the inferred decline of apparent temperature sensitivity of leaf unfolding over the last three decades (Extended Data Fig. 3a–c).

Linear regression analyses (using both ordinary least squares and reduced major axis regressions) of the dates of leaf unfolding against mean air temperature over the pre-season were performed for each species at each site during the three study periods: 1980–2013 (minimum 15-year records required per site), 1980–1994, and 1999–2013 (minimum 7-year records required per site, satisfied simultaneously for the two latter study periods). Similar results, that is, significant decreases in S_T , were observed using reduced major axis regression method (Extended Data Fig. 5b), we therefore only present the results using ordinary least squares method. The regression coefficient was defined as the apparent temperature sensitivity of leaf unfolding (S_T) that reflects the change in leaf unfolding date per unit increase in mean temperature during the pre-season. This is not the ‘actual’ physiological sensitivity to temperature, given that other climate-related variables, such as

chilling, photoperiod, solar radiation and precipitation, also co-determine the leaf unfolding process and determine the emerging S_T value diagnosed from the pre-season temperature^{8,12,22}. The mean S_T across all sites was calculated for individual species and for combined totals for all species for these three periods. The frequency distributions of S_T across all species and sites for 1980–1994 and 1999–2013 were determined. The differences in mean S_T during 1999–2013 and 1980–1994 were tested using independent *t*-tests for each and across species.

To investigate the effect of the chilling requirement and variance in spring temperature on S_T , we calculated species-specific variances in spring temperature and chilling requirements at each site. The spring temperature variance was calculated as the s.d. of mean temperature during the pre-season. The chilling requirement is normally defined as the length of the period (days or hours) during which temperature remains within a specific range. Most previous studies have reported that temperatures slightly above freezing are most effective in satisfying the chilling requirement³¹ and have suggested that the temperature range between 0 °C and 5 °C is the most effective across species. To calculate the chilling requirement, we therefore summed the days when daily temperature was within this range:

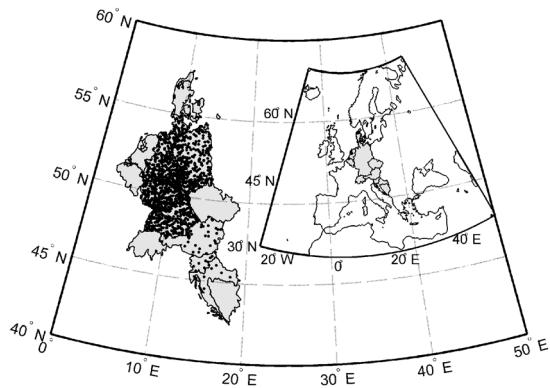
$$CD_{\text{req}}(t) = \sum_{t_0}^{t_{LF}} 1 \quad \text{if } 0 \leq T_t \leq 5$$

where CD_{req} is the chilling requirement, t_{LF} is the day of leaf unfolding, T_t is the daily mean temperature on day t , and t_0 is the start date for chilling accumulation. t_0 was fixed at 1 November in the year before leaf unfolding. We also tested another commonly used temperature threshold, 5 °C²⁴, and included all temperatures below this threshold.

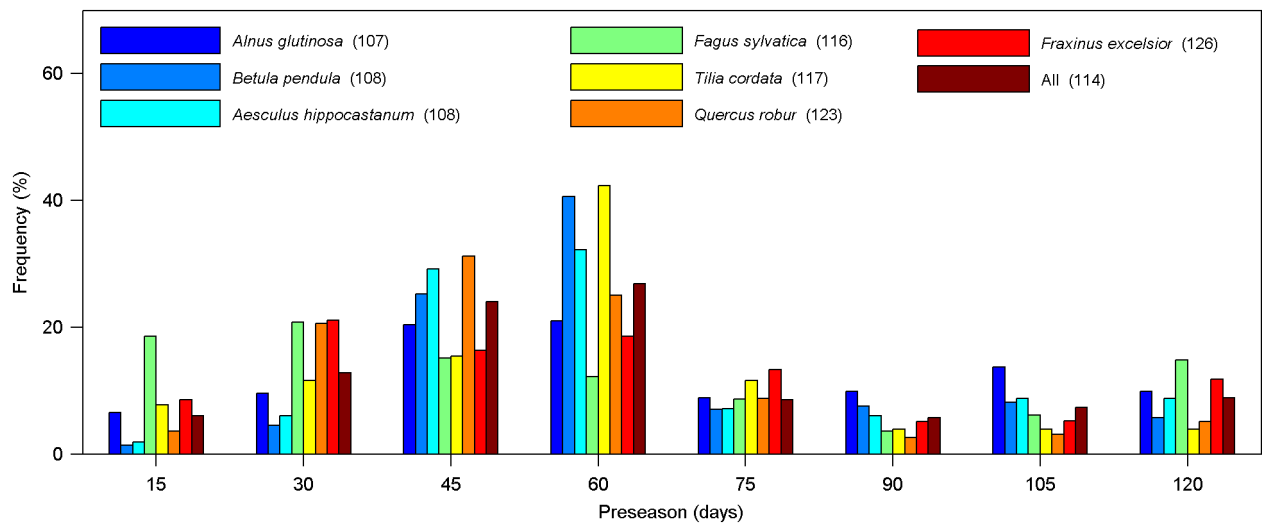
$$CD_{\text{req}}(t) = \sum_{t_0}^{t_{LF}} 1 \quad \text{if } T_t \leq 5$$

The differences in mean during 1999–2013 and 1980–1994 were tested using independent *t*-tests for each and across species. Ordinary least squares linear regression was applied to determine the temporal change in the chilling requirement for 1980–2013 and to determine the correlation between chilling accumulation and S_T .

30. Fu, Y. S. H. *et al.* Variation in leaf flushing date influences autumnal senescence and next year's flushing date in two temperate tree species. *Proc. Natl Acad. Sci. USA* **111**, 7355–7360 (2014).
31. Coville, F. V. The influence of cold in stimulating the growth of plants. *Proc. Natl Acad. Sci. USA* **6**, 434–435 (1920).

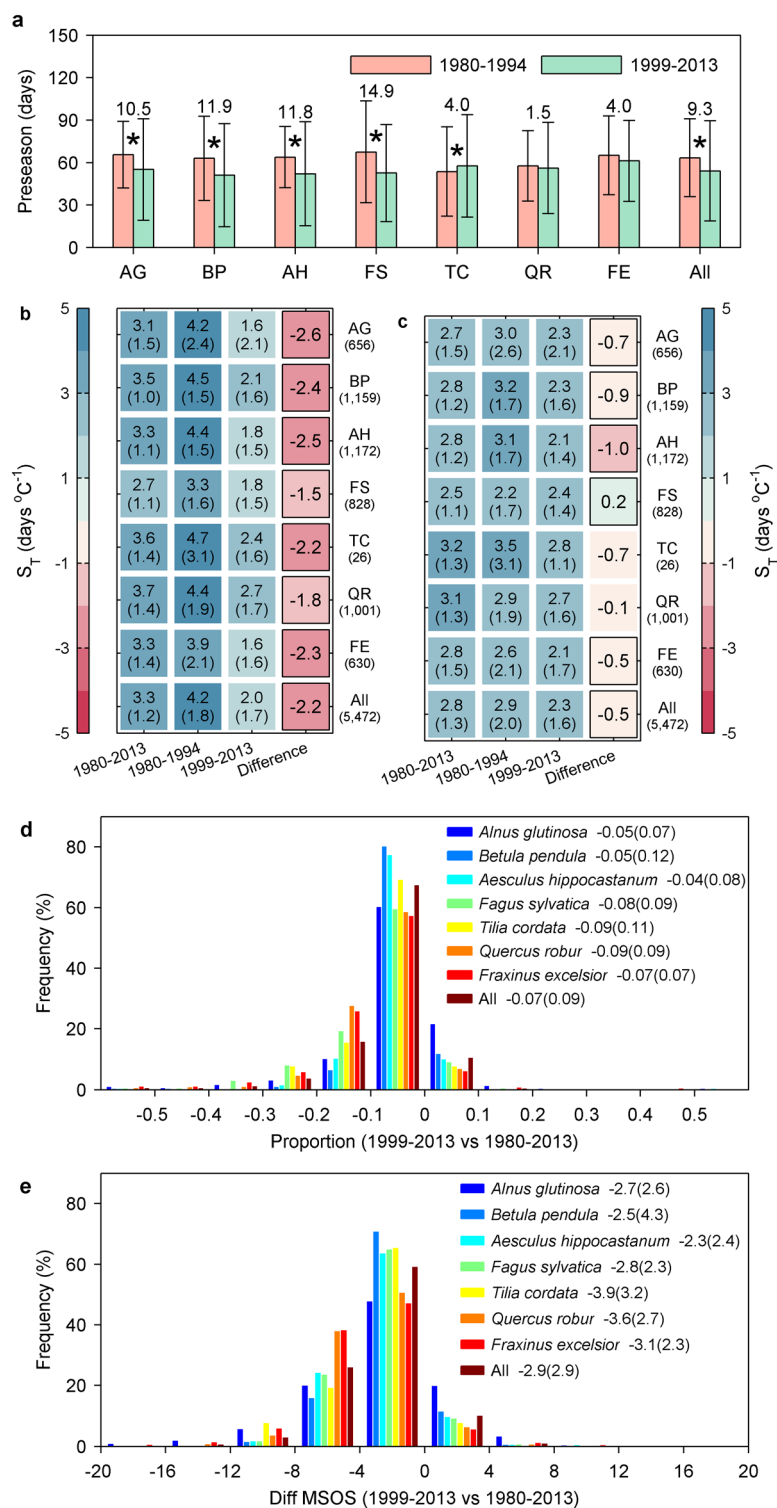


Extended Data Figure 1 | The distribution of the sites. The data were obtained from the Pan European Phenology network (<http://www.pep725.eu/>).



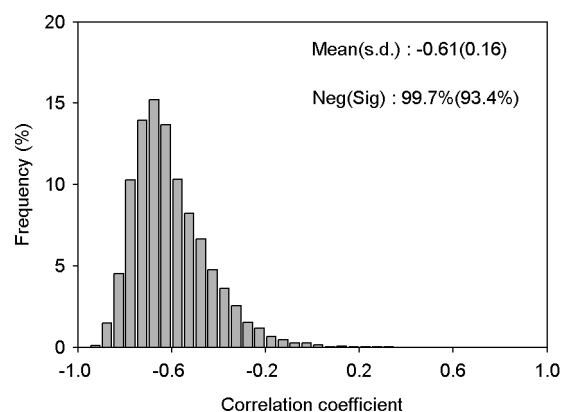
Extended Data Figure 2 | The distribution of preseason length for individual species and for combined totals for all species. The optimal preseason was defined as the period before leaf unfolding for which the

correlation coefficient between leaf unfolding and temperature was highest. The numbers in the brackets are the mean dates of leaf unfolding across all sites.

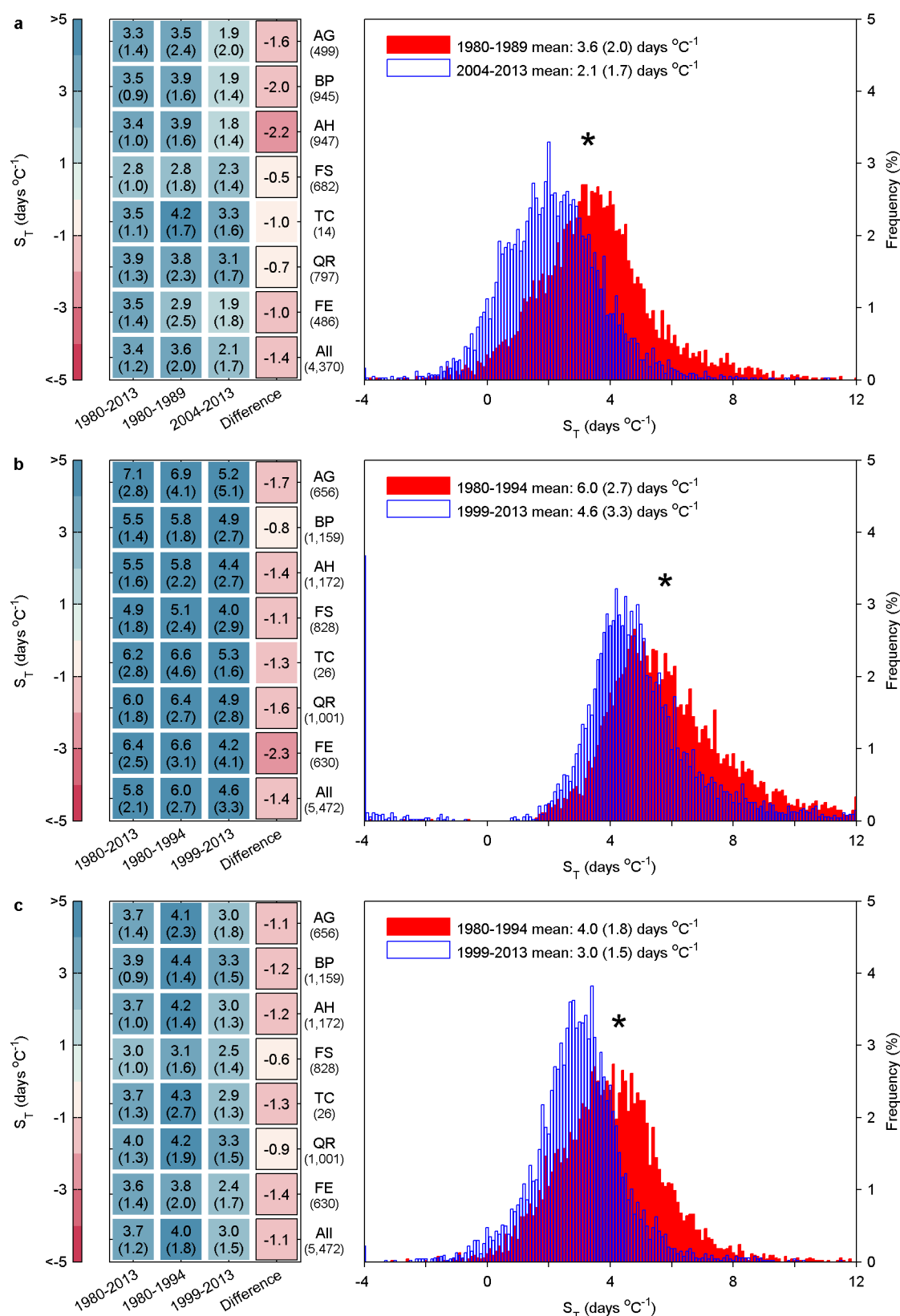


Extended Data Figure 3 | Changes of apparent temperature sensitivity of leaf unfolding between 1980–1994 and 1999–2013. **a–c,** Same as Fig. 1, but the S_T was calculated based on the preseason that was determined either in the time period 1980–1994 (**b**) or in 1999–2013 (**c**). The differences in preseason lengths are provided for individual species and for combined totals for all species (**a**), and the figures above bars are the mean absolute preseason difference between two periods. For **b** and **c**, species-specific S_T and its s.d. (in brackets) across all sites in three periods and its difference between 1999–2013 and 1980–1994. The colour scale indicates the magnitude of S_T .

The number of sites for each species are in brackets under the species name. **d, e,** The distribution of the proportion and corresponding days (**e**) of the encroachment of phenology dates into the preseason temperature that the preseason was determined on the period 1980–2013. The proportion was defined as the difference of the mean leaf unfolding dates (diff MSOS) between the period 1999–2013 and 1980–2013 (which is the end date of the preseason temperature that was used to calculate the S_T) divided by the preseason length in days. The mean values and s.d. (in brackets) are provided for individual species and for combined totals for all species.

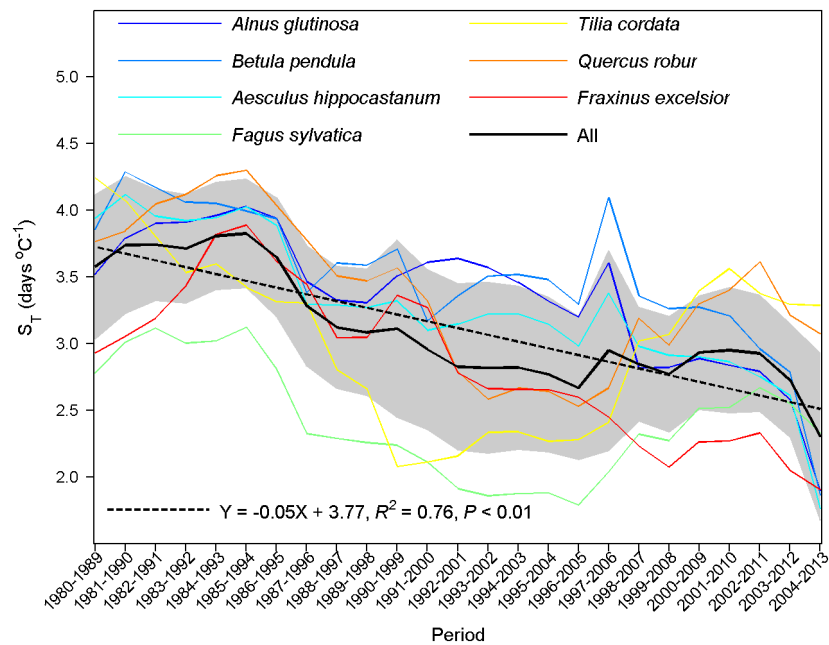


Extended Data Figure 4 | The distribution of partial correlation coefficients between preseason temperature and leaf unfolding dates over the time period 1980–2013. The mean (and s.d.) of the correlation coefficients across all species and sites are provided. The percentages of negative correlations and statistically significant negative correlations (Neg(Sig)) are also provided.



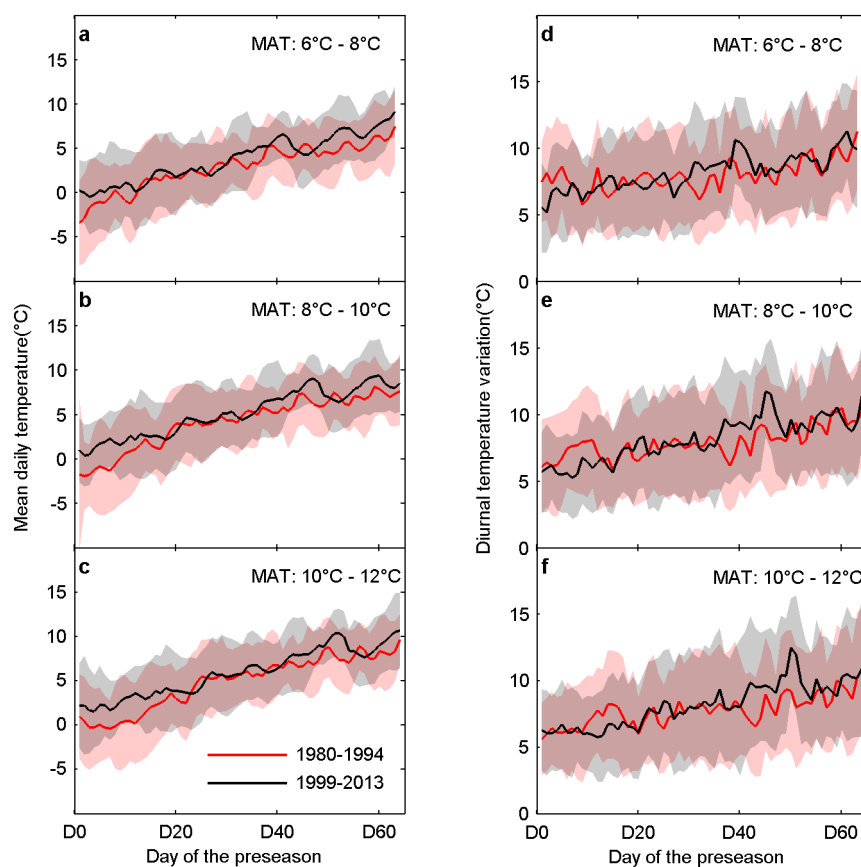
Extended Data Figure 5 | Changes of apparent temperature sensitivity of leaf unfolding determined by different methods. a–c, The S_T were analysed in two 10-year periods (a), were calculated using the reduced major axis (RMA) regression (b), or were calculated based on another climate forcing data set (CRU-NCEP v5, c). Species-specific S_T and s.d. (in brackets) across all sites in three periods and the difference between the two study periods are provided.

The colour scale indicates the magnitude of S_T . The number under the species name is the number of sites. The histograms show the distribution of S_T across all species and sites in two different periods and the mean S_T and s.d. (in brackets). The asterisk indicates a significant difference of S_T between the two periods at $P < 0.05$.



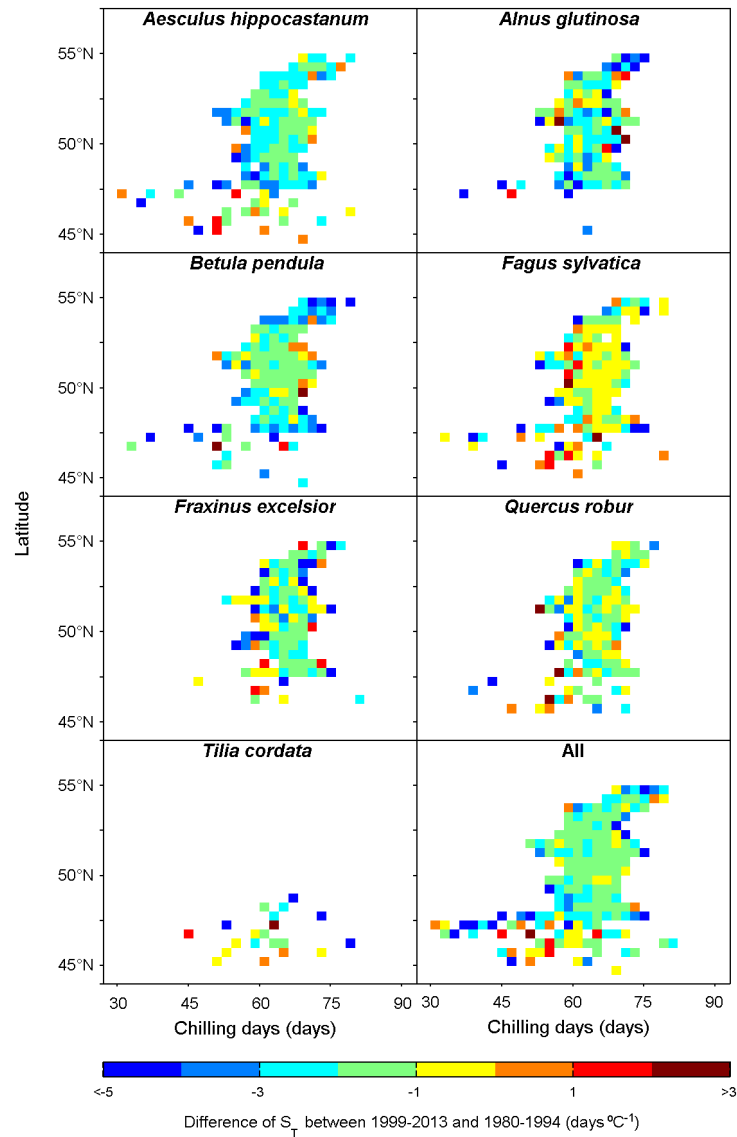
Extended Data Figure 6 | Changes of apparent temperature sensitivity of leaf unfolding over time. Same as Fig. 1c, but temporal change of S_T with 10-year moving windows from 1980 to 2013. The S_T was calculated using

simple linear regression. The black line indicates the average across all species, and the grey area indicates one s.d. either side of the mean. The dotted line indicates the linear regression.



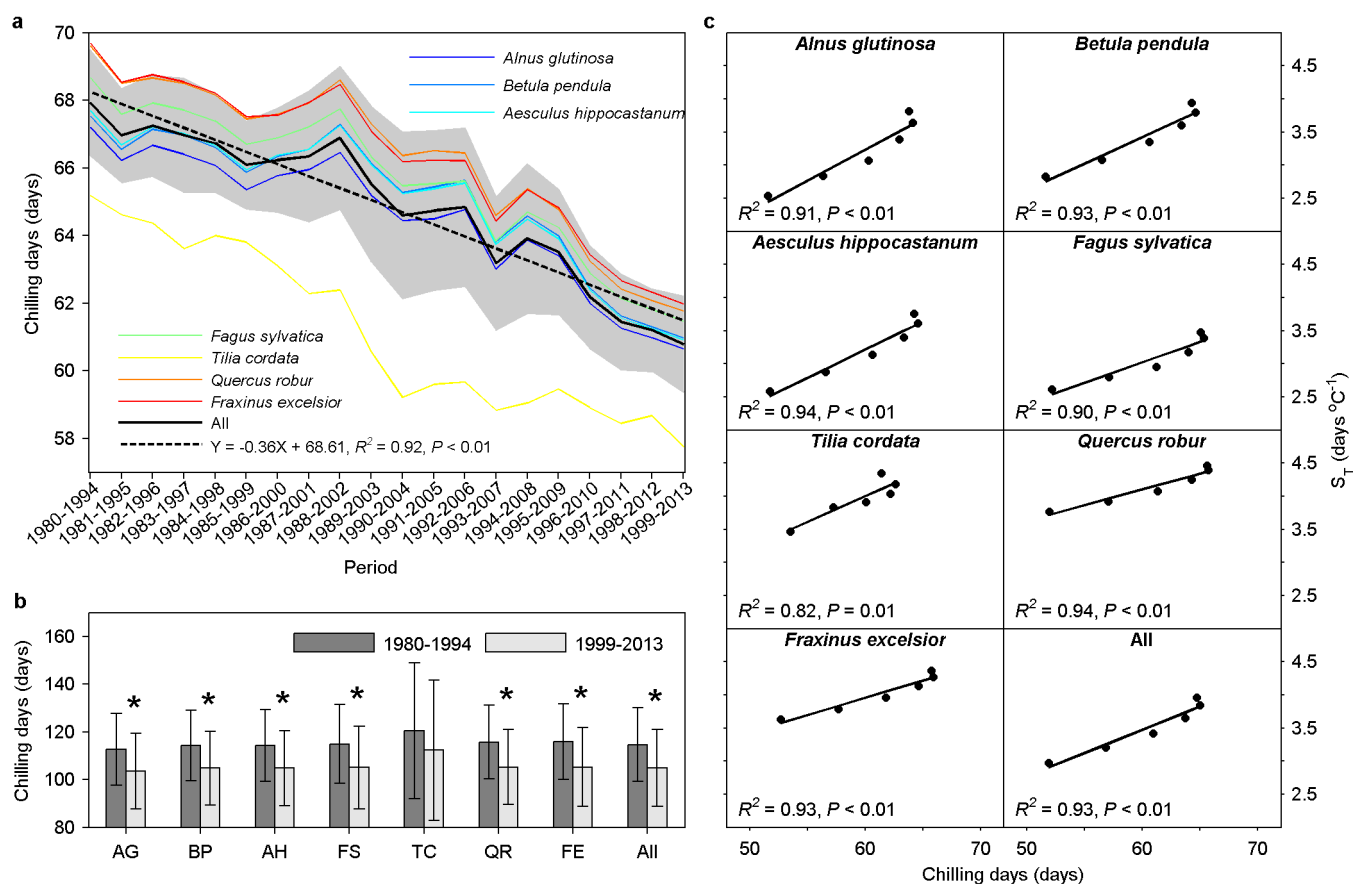
Extended Data Figure 7 | The differences in climatology over the preseason. The fluctuations in mean daily temperature (left) and diurnal variation temperature ($T_{\max} - T_{\min}$, right) over the preseason across all sites during the

time period 1980–1994 and 1999–2013 in three MAT groups, that is, (top panels) 6–8 °C, (middle panels) 8–10 °C and (bottom panels) 10–12 °C. The preseason was determined over the period 1980–2013.



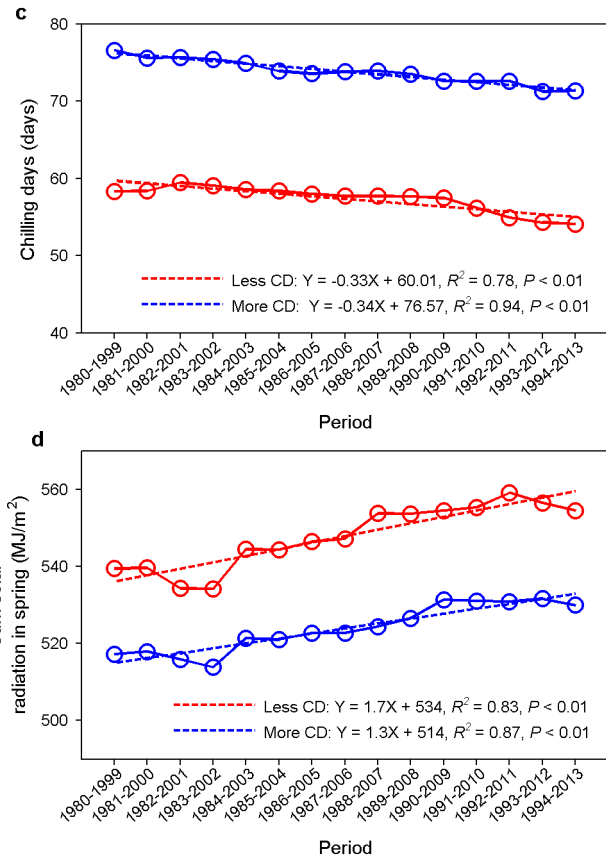
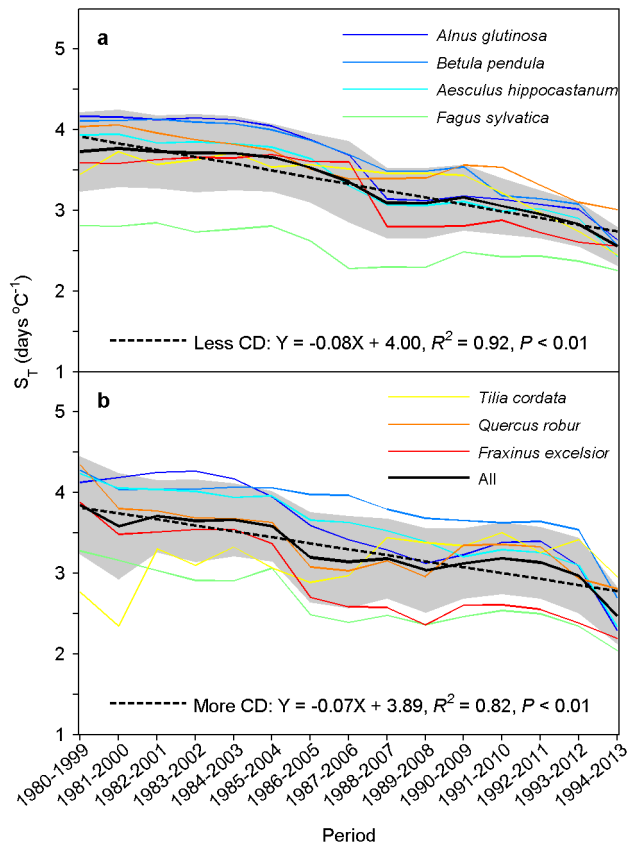
Extended Data Figure 8 | Spatial difference in apparent temperature sensitivity of leaf unfolding reduction. The difference of S_T for each species and across all species studied between two time periods, 1999–2014 and

1980–1994, at different latitudes (bin: 0.5°) and chilling conditions (bin: two chilling days). The colour scales indicate the differences of S_T between the two periods.



Extended Data Figure 9 | Changes in chilling accumulation and modelled correlation between chilling and apparent temperature sensitivity of leaf unfolding. **a**, Chilling accumulation for individual species and for combined totals for all species with 15-year moving windows from 1980 to 2013. The chilling accumulation was calculated as chilling days when daily temperature was between 0 °C and 5 °C from 1 November to the average date of leaf unfolding. The black line indicates the average across all species, and the grey area indicates one s.d. either side of the mean. The dotted line indicates the linear regression. **b**, Same as Fig. 2b, but chilling accumulation was calculated as

chilling days when daily temperature was below 5 °C from 1 November to the average date of leaf unfolding. The asterisks indicate significant differences at $P < 0.05$. **c**, The modelled (unified model) S_T under different artificial winter warming conditions. The temperature in winter, defined as the period from the 1 November to 31 January, was warmed by +1 °C to +5 °C over the period 1980–2013. The points with most chilling days indicate the real winter temperatures, and each of the other points indicate one winter warming treatment. The lines indicate simple linear regressions.



Extended Data Figure 10 | Changes in apparent temperature sensitivity of leaf unfolding between years with more or less chilling. **a, b**, S_T for years with less chilling (**a**) and more chilling (**b**) with a 20-year moving window for 1980–2013. For each 20-year series, we divided the 20 years into two groups based on the mean chilling accumulation (chilling was accumulated when daily temperature within the temperature range between 0 °C and 5 °C from 1 November to the day of leaf unfolding). The 10 years with chilling higher than the overall mean were defined as more chilling, and the other 10 years were

defined as less chilling. The black lines indicate the average across all species, and the grey area indicates one s.d. either side of the mean. The dotted lines are the linear regressions. **c**, Chilling accumulation for years with less chilling (red line) and more chilling (blue line) with a 20-year moving window for 1980–2013. **d**, The mean radiation sum over the pre-season for years with less chilling (red line) and more chilling (blue line) with a 20-year moving window for 1980–2013. The pre-season was determined over the period 1980–2013.

New genomic and fossil data illuminate the origin of enamel

Qingming Qu^{1*}, Tatjana Haitina^{1*}, Min Zhu² & Per Erik Ahlberg¹

Enamel, the hardest vertebrate tissue, covers the teeth of almost all sarcopterygians (lobe-finned bony fishes and tetrapods) as well as the scales and dermal bones of many fossil lobe-fins^{1–5}. Enamel deposition requires an organic matrix containing the unique enamel matrix proteins (EMPs) amelogenin (AMEL), enamelin (ENAM) and ameloblastin (AMBN)⁶. Chondrichthyans (cartilaginous fishes) lack both enamel and EMP genes^{7,8}. Many fossil and a few living non-teleost actinopterygians (ray-finned bony fishes) such as the gar, *Lepisosteus*, have scales and dermal bones covered with a proposed enamel homologue called ganoine^{1,9}. However, no gene or transcript data for EMPs have been described from actinopterygians^{10,11}. Here we show that *Psarolepis romeri*, a bony fish from the Early Devonian period, combines enamel-covered dermal odontodes on scales and skull bones with teeth of naked dentine, and that *Lepisosteus oculatus* (the spotted gar) has *enam* and *ambn* genes that are expressed in the skin, probably associated with ganoine formation. The genetic evidence strengthens the hypothesis that ganoine is homologous with enamel. The fossil evidence, further supported by the Silurian bony fish *Andreolepis*, which has enamel-covered scales but teeth and odontodes on its dermal bones made of naked dentine^{12–16}, indicates that this tissue originated on the dermal skeleton, probably on the scales. It subsequently underwent heterotopic expansion across two highly conserved patterning boundaries (scales/head–shoulder and dermal/oral) within the odontode skeleton.

Most vertebrates possess skeletal structures containing one or more of the tissues dentine, enamel and enameloid; we refer to the totality of such structures as the ‘odontode skeleton’ of the animal. The genetic basis of these tissues is currently the subject of intensive research, and enamel formation in tetrapods is already quite well understood. During the secretory stage, the ameloblasts produce the EMPs AMEL, ENAM and AMBN, which are progressively degraded during the following maturation stage, when the secreted matrix is replaced by the mineral hydroxyapatite. During enamel maturation, the ameloblasts produce the proteins amelotin (AMTN), odontogenic ameloblast-associated (ODAM) and secretory calcium-binding phosphoprotein-proline-glutamine-rich 1 (SCPPPQ1); AMTN is important for the mineralization process in mammals^{17,18}. EMPs and maturation-stage proteins belong to the proline-glutamine (P/Q)-rich secretory calcium-binding phosphoprotein (SCPP) family. With the exception of AMEL, these genes are ancestrally located in one cluster, together with acidic bone and dentine SCPP genes^{10,11}.

This rich molecular data set creates a new genomic tool for investigating the evolution of enamel. However, tetrapods are highly derived in having only an oral odontode skeleton (Fig. 1). Gnathostomes primitively also have an extensive dermal (that is, external) odontode skeleton, which is retained in living chondrichthyans, in the sarcopterygian fish *Latimeria*, and in bichirs (*Polypterus*, *Erpetoichthys*) andgars (*Lepisosteus*, *Atractosteus*) among actinopterygians^{1,9}. The oral and dermal odontode skeletons have been argued to be fundamentally

separate systems, with the epithelial component of the odontode bud of endodermal origin in the one but of ectodermal origin in the other¹⁹. Putative homologues of tooth enamel occur on the dermal odontode skeletons of both sarcopterygians and actinopterygians. By combining new genomic and palaeontological data, we are able to present a well-supported hypothesis for the origin, distribution and patterning boundaries of this tissue.

It has long been argued from histological criteria that all sarcopterygians carry true enamel on their teeth and (if present) dermal odontode skeleton^{2–5}. This is confirmed by the presence of AMEL, ENAM, AMBN, AMTN, ODA and SCPPPQ1, not only in toothed tetrapods but also in the coelacanth *Latimeria chalumnae*¹⁰ (Fig. 2 and Extended Data Fig. 1). As *Latimeria* and tetrapods bracket the entire sarcopterygian crown group, the genes can be inferred to have been present even in the earliest fossil crown-group sarcopterygians, such as *Youngolepis*² (Fig. 1).

Outside the Sarcopterygii the situation is less clear cut (Fig. 1). In early fossil actinopterygians, and in the living bichirs andgars, the dermal skeleton is covered in ganoine, a tissue resembling enamel^{1,20} but sometimes identified as enameloid in older literature²¹. The teeth of crown-group actinopterygians have tips covered with acrodin or cap enameloid²², but this appears to be absent in stem actinopterygians such as *Cheirolepis*⁵. Primitive crown-group actinopterygians, including bichirs andgars, have a thin layer of ‘collar ganoine’ or ‘collar enamel’ on the tooth shaft below the acrodin cap⁹. Teleosts (advanced actinopterygians) never have ganoine on the dermal skeleton, and the majority also lack collar enamel.

The EMP genes *enam*, *ambn* and *amel* are absent in sequenced teleosts, and of the maturation-stage proteins only Odam is present¹¹. This means that acrodin is not a product of EMPs and that its mineralization programme differs from that of tetrapod enamel. However, because the sequenced teleosts lack not only EMP genes but also collar enamel and ganoine, the published genomic data fail to illuminate the molecular basis and homology of these actinopterygian tissues. To remedy the problem, we searched for EMP genes in the recently sequenced genome of the spotted gar, *L. oculatus*²³ (Extended Data Table 1 and Supplementary Information). On the basis of synteny and gene structure analysis, we have identified in the gar genome a cluster of P/Q-rich SCPP genes related to hard tissues. This cluster, located on chromosome LG4, can be divided into three main groups: EMP genes, represented by *ambn* and *enam*; maturation-stage genes, represented by orthologues of tetrapod AMTN and SCPPPQ1, and teleost odam; and actinopterygian-specific SCPP genes, previously reported only in teleosts (Fig. 2 and Extended Data Fig. 1). Gar *ambn* and *enam* were identified as RNA-sequencing (RNA-seq) transcripts in the skin (Extended Data Table 1), suggesting that they have a role in ganoine deposition and thus that there is primary molecular homology between ganoine matrix and enamel matrix. However, in gar, *amel* is not found in its usual position (Fig. 2) and may be absent from the genome; this gene codes for the major protein in the extracellular matrix of developing enamel of

¹Subdepartment of Evolution and Development, Department of Organismal Biology, Uppsala University, Norbyvägen 18A, SE 75236 Uppsala, Sweden. ²Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China.

*These authors contributed equally to this work.

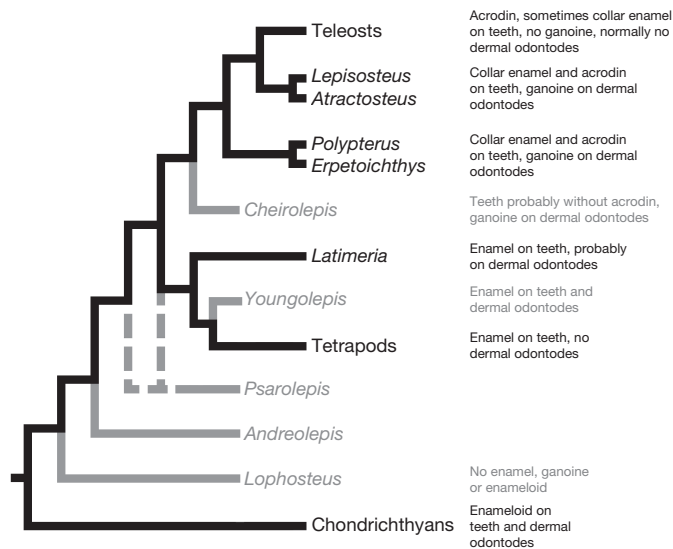


Figure 1 | Phylogeny of odontode surface tissues. Simplified phylogeny of gnathostomes (topology from refs 27, 28) showing published data on the distribution of surface tissues on the odontode skeleton. Black branches, names and text indicate extant taxa; grey branches, names and text indicate fossil taxa. The surface tissues of *Psarolepis* and *Andreolepis* are discussed in the text.

tetrapod teeth^{6,10}, suggesting that sarcopterygian enamel differs from actinopterygian ganoin in this respect.

These results strongly support the histologically based argument^{1,20} for primary homology between ganoin and enamel, and indicate that tooth enamel, dermal enamel/ganoin, and enamel-generating genes were all present at the osteichthyan crown-group node. Furthermore, as the molecular data support the validity of the histological criteria used to identify enamel^{1,20}, they indirectly support the histological identification of this tissue in fossils. By contrast, chondrichthyans appear to lack EMP genes⁷, indicating that their enameloid matrix is non-homologous with enamel matrix.

As non-osteichthyans lack enamel¹, the tissue must have evolved in the osteichthyan stem group and is thus considerably younger than the distinction between the oral and dermal odontode skeletons. This implies that enamel became established as a tissue spanning major

pre-existing patterning boundaries, either at its time of origin or later through heterotopic expansion across patterning regions. The evolution of these patterning relationships is documented by early fossil osteichthyans. Histological data are available for two Late Silurian period stem osteichthyans, *Lophosteus* and *Andreolepis*^{14,15,24,25} (Fig. 1). *Lophosteus* lacks enamel²⁴. In *Andreolepis*, enamel is present on the scales but not on the dermal bones or teeth^{14–16,25}. However, the *Andreolepis* material lacks direct single-specimen association between skull bones and scales^{12,14}. By contrast, the possible stem osteichthyan²⁶ (or stem sarcopterygian^{13,26,27}) *Psarolepis romeri*, from the Silurian to earliest Devonian of China (Figs 1 and 3), is known from partly articulated material, providing direct associations of teeth and dermal odontodes²⁶. Sections from a parietal shield with articulated premaxillae (Fig. 3a–d and Extended Data Figs 2–4) and a lower jaw (Fig. 3e–g and Extended Data Fig. 5) confirm previous descriptions of enamel-covered dermal odontodes²⁸. The invaginated dermal process in the anterior nostril carries tooth-like denticles that are covered with enamel (Fig. 3b), and enamel is also present on the scales²⁸.

Remarkably, the only odontodes in *Psarolepis* that lack enamel are the teeth (Fig. 3c, d and Extended Data Fig. 4). In the lower jaw, a line of tiny tooth-shaped denticles marks the transition between the tooth row and the labial surface of the bone (Extended Data Fig. 5); these denticles carry enamel, whereas the teeth do not (Fig. 3f). Enamel is also absent on the coronoid teeth of the lower jaw (Fig. 3g). The absence of enamel on the teeth is not a preservational artefact, as there is no evidence of abrasion or chemical erosion. Examination by scanning electron microscopy (SEM) of the microstructure of the cranial dermal enamel layer shows characteristic incremental lines and crystals normal to the surface, like the enamel on *Psarolepis* scales²⁸ (Fig. 3h, i) and on the teeth of crown-group sarcopterygians^{2,5}.

Psarolepis demonstrates that odontodes with and without enamel can coexist in one animal. This in turn means that the attribution of dispersed osteichthyan-like scales and head bones from the Late Silurian of Gotland to the single taxon *Andreolepis hedei*^{12,14} is not contradicted by the presence of enamel only on the scales; *Andreolepis* can be interpreted with some confidence as a single taxon with enamel-covered scales, rather than a chimera of two taxa. Together, the positive molecular identification of ganoin as enamel and the recognition of restricted enamel distribution in *Andreolepis* and *Psarolepis* provide important evidence for the evolution of this tissue (Fig. 4).

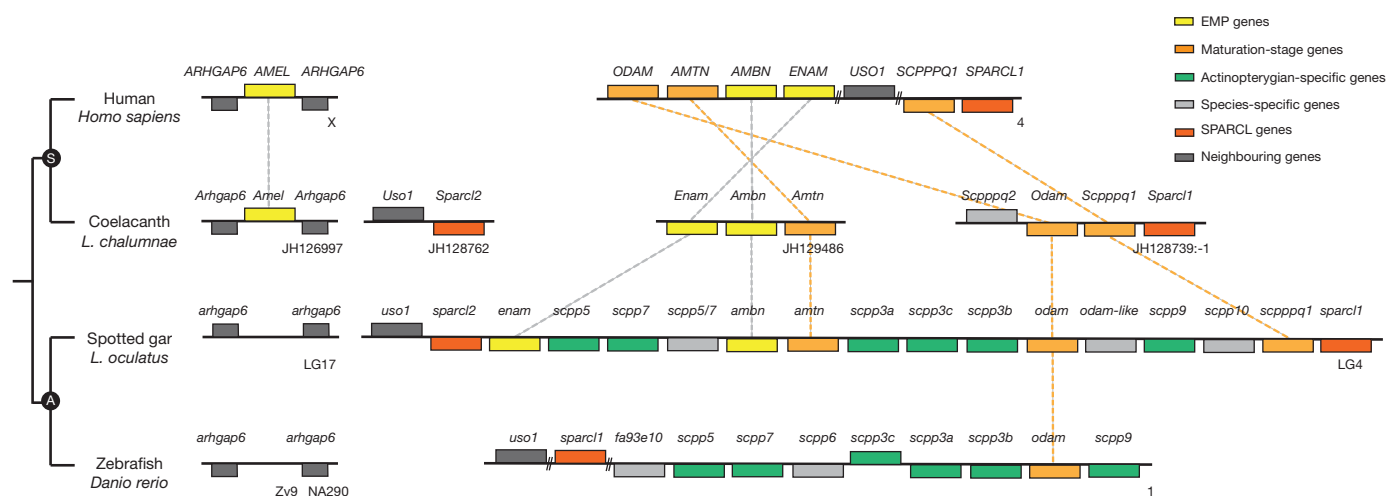


Figure 2 | P/Q-rich SPP gene cluster. The arrangement and direction of the P/Q-rich SPP genes on chromosomes of human (*Homo sapiens*), spotted gar (*Lepisosteus oculatus*) and zebrafish (*Danio rerio*); and genomic scaffolds of coelacanth (*Latimeria chalumnae*). Human P/Q-rich SPP genes lacking orthologues in coelacanth, gar and zebrafish genomes are not shown. 'S' indicates sarcopterygian group and 'A' indicates actinopterygian group. Yellow boxes, enamel matrix protein genes; orange boxes, maturation-stage genes; green boxes, actinopterygian-specific P/Q-rich SPP genes; light grey boxes, species-specific SPP genes; red boxes, SPARCL genes; dark grey boxes, other neighbouring genes.

indicates sarcopterygian group and 'A' indicates actinopterygian group. Yellow boxes, enamel matrix protein genes; orange boxes, maturation-stage genes; green boxes, actinopterygian-specific P/Q-rich SPP genes; light grey boxes, species-specific SPP genes; red boxes, SPARCL genes; dark grey boxes, other neighbouring genes.

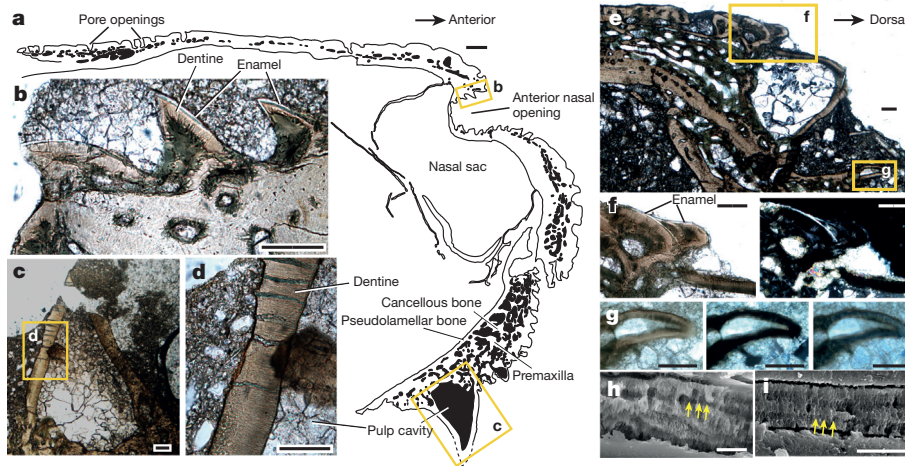


Figure 3 | Sagittal section (IVPP V17756.3) of the cranial roof and transverse section (IVPP V19360.1) of the lower jaw of *P. romeri*. **a**, Outline of thin section showing positions of **b** and **c**. **b**, Enamel-covered denticles on the invaginated dermal process in the nostril, under transmission light. **c**, **d**, A large tooth made of dentine without hypermineralized surface tissue (**c**), and detail of same (**d**), under transmission light. **e**, Cross-section of lower jaw showing, from top left to bottom right, cosmine (fused odontodes enclosing a pore-canal system²⁹) covering labial surface, teeth with large pulp cavities on

dentary, and smaller teeth on coronoid. **f**, Detail of **e** showing that the labial surface is covered by an enamel layer, whereas the immediate adjacent tooth lacks enamel, under transmission light (left) and cross-polarized light (right). **g**, Detail of **e** showing a coronoid tooth that is lacking enamel, under transmission light (left), cross-polarized light (middle) and Nomarski interference (right). **h**, SEM of enamel layer on the skull roof (IVPP V17756). **i**, SEM of enamel layer on the scale (IVPP V17758.2). Scale bars, 300 μ m (**a**), 100 μ m (**b–g**), 5 μ m (**h–i**).

The fossils reveal two patterning boundaries affecting the distribution of enamel, one between the dermal and oral systems (*Psarolepis*) and one within the dermal system, between the scales and skull bones (*Andreolepis*). The dermal/oral boundary has been discussed extensively in relation to the evolution of the odontode skeleton¹⁹, but the boundary between the scales and the bones of the shoulder girdle and head has received much less consideration, even though it is present in placoderms²⁷ and thus apparently primitive for jawed vertebrates. This is probably because it has been lost in chondrichthyans, frequently (but inappropriately) used as model primitive

gnathostomes (Fig. 4b). We hypothesize that enamel originated on the scales, before colonizing the dermal bones and finally the teeth. Ganoine-covered actinopterygians, including the extant gars and bichirs, consistently show sharp morphological boundaries between the odontode skeletons of the scales, dermal skull bones and teeth, indicating that both patterning boundaries remain active. We predict that the presence of these boundaries, and the stepwise extension of enamel across them from an origin on the scales, will be reflected in the regulatory network architecture of the EMP and other enamel-related genes in gar and bichir. By contrast, tetrapods will have a simplified

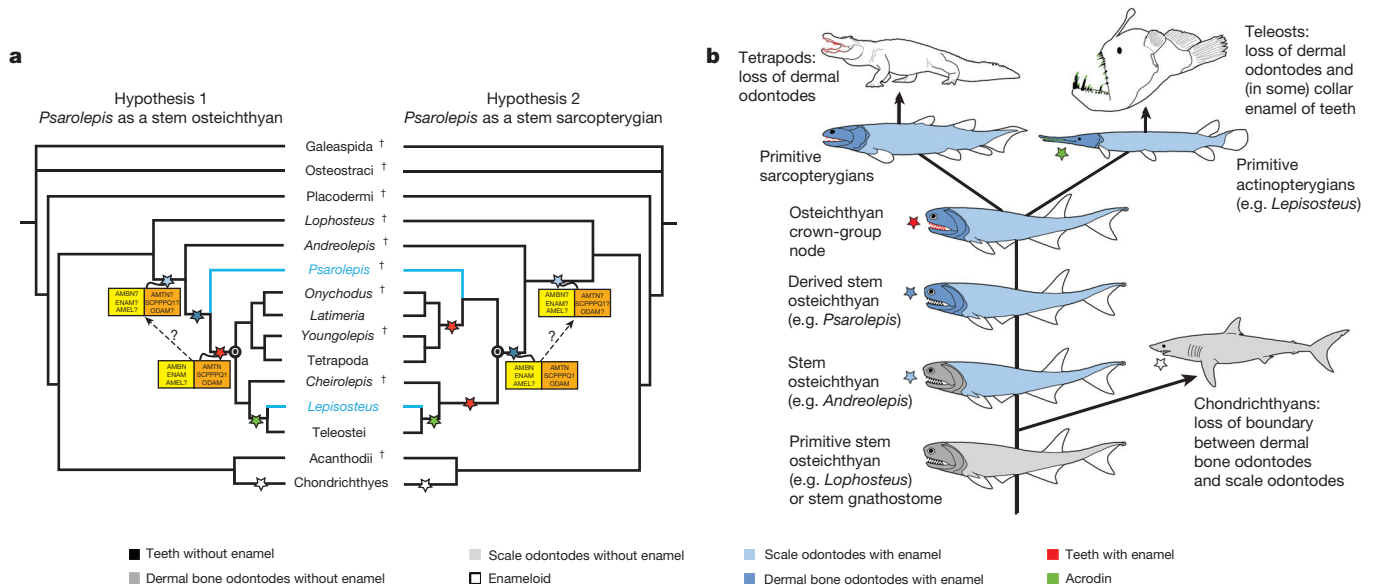


Figure 4 | Scenarios about the early evolution of enamel. **a**, Alternative phylogenetic placements of *P. romeri* showing implications for the evolution of enamel. Topologies are based on refs 27, 28. Daggers denote extinct taxa; 'O', the osteichthyan crown-group node; yellow boxes, EMPs; orange boxes, maturation-stage proteins. **b**, Evolutionary scenario based on hypothesis 1 in **a**. Stem osteichthyans and crown-group node osteichthyan represented by a schematic fish modelled on the Devonian stem actinopterygian *Cheirolepis trailli*; primitive

actinopterygians represented by the spotted gar, *L. oculatus*; primitive sarcopterygians represented by the Devonian tetrapodomorph *Osteolepis macrolepidotus*; tetrapods represented by an alligator, *Alligator mississippiensis*; teleosts represented by a deep-sea anglerfish, *Melanocetus johnsoni*; chondrichthyans represented by a mako shark, *Isurus oxyrinchus*. All drawings are original except *Cheirolepis* and *Osteolepis*, which are modified from refs 29, 30 with permission. **a**, **b**, Coloured stars indicate first appearance of a tissue.

regulatory network in which components driving expression in the dermal skeleton have been lost.

The exact evolutionary importance of the absence of tooth enamel in *Psarolepis* depends on the phylogenetic position of the taxon: if it is a stem osteichthyan²⁶, tooth enamel is a synapomorphy of the osteichthyan crown group, but if it is a stem sarcopterygian^{13,27}, tooth enamel either evolved independently in sarcopterygians and actinopterygians, or has been lost in *Psarolepis* (Fig. 4a). Loss seems unlikely for functional reasons. However, a more detailed phylogenetic analysis of the earliest osteichthyans¹³, together with detailed comparative investigation of the molecular regulatory networks of sarcopterygians and actinopterygians, will be required to clarify exactly when and how enamel colonized the teeth.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 April; accepted 28 July 2015.

Published online 23 September 2015.

- Sire, J.-Y., Donoghue, P. C. J. & Vickaryous, M. K. Origin and evolution of the integumentary skeleton in non-tetrapod vertebrates. *J. Anat.* **214**, 409–440 (2009).
- Chang, M.-M. & Smith, M. M. Is *Youngolepis* a porolepiform? *J. Vertebr. Paleontol.* **12**, 294–312 (1992).
- Smith, M. M. Enamel in the oral teeth of *Latimeria chalumnae* (Pisces: Actinistia): a scanning electron microscope study. *J. Zool.* **185**, 355–369 (1978).
- Satchell, P. G., Shuler, C. F. & Diekwisch, T. G. H. True enamel covering in teeth of the Australian lungfish *Neoceratodus forsteri*. *Cell Tissue Res.* **299**, 27–37 (2000).
- Smith, M. M. in *Structure, Function and Evolution of Teeth* (eds Smith, P. & Tchernov, E.) 73–101 (Freund, 1992).
- Sire, J.-Y., Davit-Béal, T., Delgado, S. & Gu, X. The origin and evolution of enamel mineralization genes. *Cells Tissues Organs* **186**, 25–48 (2007).
- Venkatesh, B. et al. Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**, 174–179 (2014).
- Grady, J. E. Tooth development in sharks. *Arch. Oral Biol.* **15**, 613–619 (1970).
- Sasagawa, I., Ishiyama, M., Yokosuka, H. & Mikami, M. Teeth and ganoid scales in *Polypterus* and *Lepisosteus*, the basic actinopterygian fish: an approach to understand the origin of the tooth enamel. *J. Oral Biosciences* **55**, 76–84 (2013).
- Kawasaki, K. & Amemiya, C. T. SCPP genes in the coelacanth: tissue mineralization genes shared by sarcopterygians. *J. Exp. Zool. B Mol. Dev. Evol.* **322**, 390–402 (2014).
- Kawasaki, K. The SCPP gene family and the complexity of hard tissues in vertebrates. *Cells Tissues Organs* **194**, 108–112 (2011).
- Botella, H., Blom, H., Dorka, M., Ahlberg, P. E. & Janvier, P. Jaws and teeth of the earliest bony fishes. *Nature* **448**, 583–586 (2007).
- Zhu, M. et al. The oldest articulated osteichthyan reveals mosaic gnathostome characters. *Nature* **458**, 469–474 (2009).
- Gross, W. Fragliche Actinopterygier-Schuppen aus dem Silur Gotlands. *Lethaia* **1**, 184–218 (1968).
- Qu, Q., Sanchez, S., Blom, H., Tafforeau, P. & Ahlberg, P. E. Scales and tooth whorls of ancient fishes challenge distinction between external and oral 'teeth'. *PLoS One* **8**, e71890 (2013).
- Cunningham, J. A., Rücklin, M., Blom, H., Botella, H. & Donoghue, P. C. J. Testing models of dental development in the earliest bony vertebrates, *Andreolepis* and *Lophosteus*. *Biol. Lett.* **8**, 833–837 (2012).
- Moffatt, P., Wazen, R. M., Dos Santos Neves, J. & Nanci, A. Characterisation of secretory calcium-binding phosphoprotein-proline-glutamine-rich 1: a novel basal lamina component expressed at cell-tooth interfaces. *Cell Tissue Res.* **358**, 843–855 (2014).
- Nakayama, Y., Holcroft, J. & Ganss, B. Enamel hypomineralization and structural defects in amelotin-deficient mice. *J. Dent. Res.* **94**, 697–705 (2015).
- Fraser, G. J. & Smith, M. M. Evolution of developmental pattern for vertebrate dentitions: an oro-pharyngeal specific mechanism. *J. Exp. Zool. B Mol. Dev. Evol.* **316B**, 99–112 (2011).
- Sire, J.-Y., Géraudie, J., Meunier, F. J. & Zylberberg, L. On the origin of ganoine: histological and ultrastructural data on the experimental regeneration of the scales of *Calamoichthys calabaricus* (Osteichthyes, Brachyopterygii, Polypteridae). *Am. J. Anat.* **180**, 391–402 (1987).
- Thomson, K. S. & McCune, A. Development of the scales in *Lepisosteus* as a model for scale formation in fossil fishes. *Zool. J. Linn. Soc.* **82**, 73–86 (1984).
- Poole, D. F. G. in *Structural and Chemical Organization of Teeth* Vol. 1 (ed. Miles, A. E. W.) 111–149 (Academic, 1967).
- Braasch, I. et al. A new model army: emerging fish models to study the genomics of vertebrate Evo-Devo. *J. Exp. Zool. B Mol. Dev. Evol.* **324**, 316–341 (2015).
- Gross, W. *Lophosteus superbus* Pander: Zähne, Zahnknochen und besondere Schuppenformen. *Lethaia* **4**, 131–152 (1971).
- Richter, M. & Smith, M. M. A microstructural study of the ganoine tissue of selected lower vertebrates. *Zool. J. Linn. Soc.* **114**, 173–212 (1995).
- Zhu, M. et al. A primitive fossil fish sheds light on the origin of bony fishes. *Nature* **397**, 607–610 (1999).
- Zhu, M. et al. A Silurian placoderm with osteichthyan-like marginal jaw bones. *Nature* **502**, 188–193 (2013).
- Qu, Q., Zhu, M. & Wang, W. Scales and dermal skeletal histology of an early bony fish *Psarolepis romeri* and their bearing on the evolution of rhombic scales and hard tissues. *PLoS ONE* **8**, e61485 (2013).
- Pearson, D. M. & Westoll, T. S. The Devonian actinopterygian *Cheirolepis* Agassiz. *Trans. R. Soc. Edinb.* **70**, 337–399 (1979).
- Jarvik, E. *Basic Structure and Evolution of Vertebrates* Vol. 1 (Academic, 1980).

Supplementary Information is available in the online version of the paper.

Acknowledgements This project was inspired in part by initial discussions with K. Kawasaki, whom we gratefully acknowledge. We thank the Broad Institute Genomics Platform and Vertebrate Genome Biology group, Spotted Gar Genome Consortium and K. Lindblad-Toh for making the data for *L. oculatus* available. We thank W. Zhang and S. Zhang for technical help with thin sections and SEM. The work was supported by the Knut and Alice Wallenberg Foundation through a Wallenberg Scholarship awarded to P.E.A., and by Vetenskapsrådet (Swedish Research Council) through a Young Researcher Grant awarded to T.H. M.Z. was funded by the National Basic Research Programme of China (2012CB821902).

Author Contributions Q.Q. and T.H. initiated the project. Q.Q. collected and analysed the palaeontological data, and produced Fig. 3 and Extended Data Figs 2–5. T.H. collected and analysed the genomic data, and produced Figs 2, 4a, Extended Data Fig. 1, Extended Data Table 1 and Supplementary Information. M.Z. provided material of *Psarolepis*. P.E.A. led the writing process, and produced Figs 1 and 4b. All authors participated in the writing process.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.E.A. (per.ahlberg@ebc.uu.se).

METHODS

Materials and thin sections. The fossils of *Psarolepis* were found in the Early Devonian Xitun Formation, Qujing, East Yunnan, China. Locality information can be found elsewhere²⁶. Thin sections were made as previously described²⁸. Microscope imaging and SEM examination are as described previously²⁸. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Identification of P/Q-rich SCPP genes in the genome of spotted gar. We searched for the P/Q-rich genes belonging to the secretory calcium-binding phosphoprotein (SCPP) family in the spotted gar (*Lepisosteus oculatus*) genome assembly LepOcu1 with incorporated RNA-seq models available in the Ensembl 77 database³¹. The region of interest has been defined as falling between the *uso1* and *sparcl1* genes on chromosome LG4 and in the region of the *arhgap6* gene on chromosome LG17. We analysed the RNA-seq models from brain, embryo, eye, heart, kidney, larvae, liver, muscle, skin and testis represented in the data set. Exon–intron organization of the transcripts was reviewed in Ensembl. Predicted protein sequences were aligned with Clustal Omega³² together with previously

reported SCPP proteins from the coelacanth (*Latimeria chalumnae*)¹⁰, zebrafish (*Danio rerio*)³³ and fugu (*Takifugu rubripes*)³⁴. Signal peptide sequences were predicted with SignalP 4.1 (ref. 35). N-linked glycosylation sites were predicted with NetNGlyc 1.0 (<http://www.cbs.dtu.dk/services/NetNGlyc/>). We carried out a careful examination of predictions regarding exon–intron boundaries, signal peptide motifs and amino acid composition of translated exons. This information is discussed in Supplementary Information and is presented in table form in Extended Data Table 1.

31. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).
32. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
33. Kawasaki, K. The SCPP gene repertoire in bony vertebrates and graded differences in mineralized tissues. *Dev. Genes Evol.* **219**, 147–157 (2009).
34. Kawasaki, K., Suzuki, T. & Weiss, K. M. Phenogenetic drift in evolution: the changing genetic basis of vertebrate teeth. *Proc. Natl Acad. Sci. USA* **102**, 18063–18068 (2005).
35. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**, 785–786 (2011).

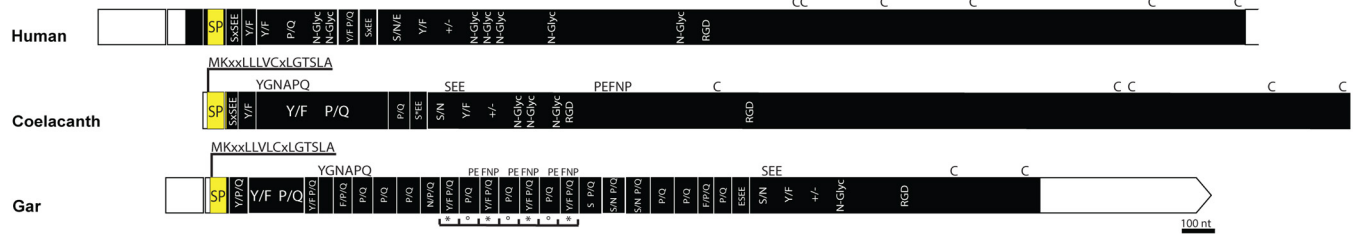
AMBN



SCPP6



ENAM



ODAM



ODAM-like



AMTN

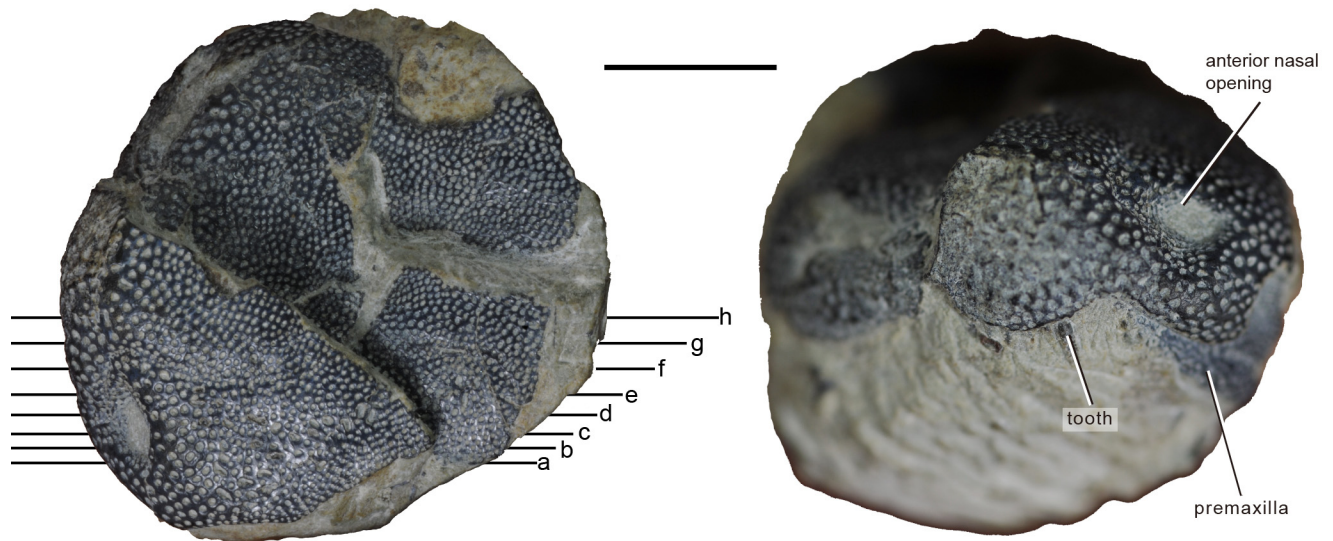


SCPPPQ1



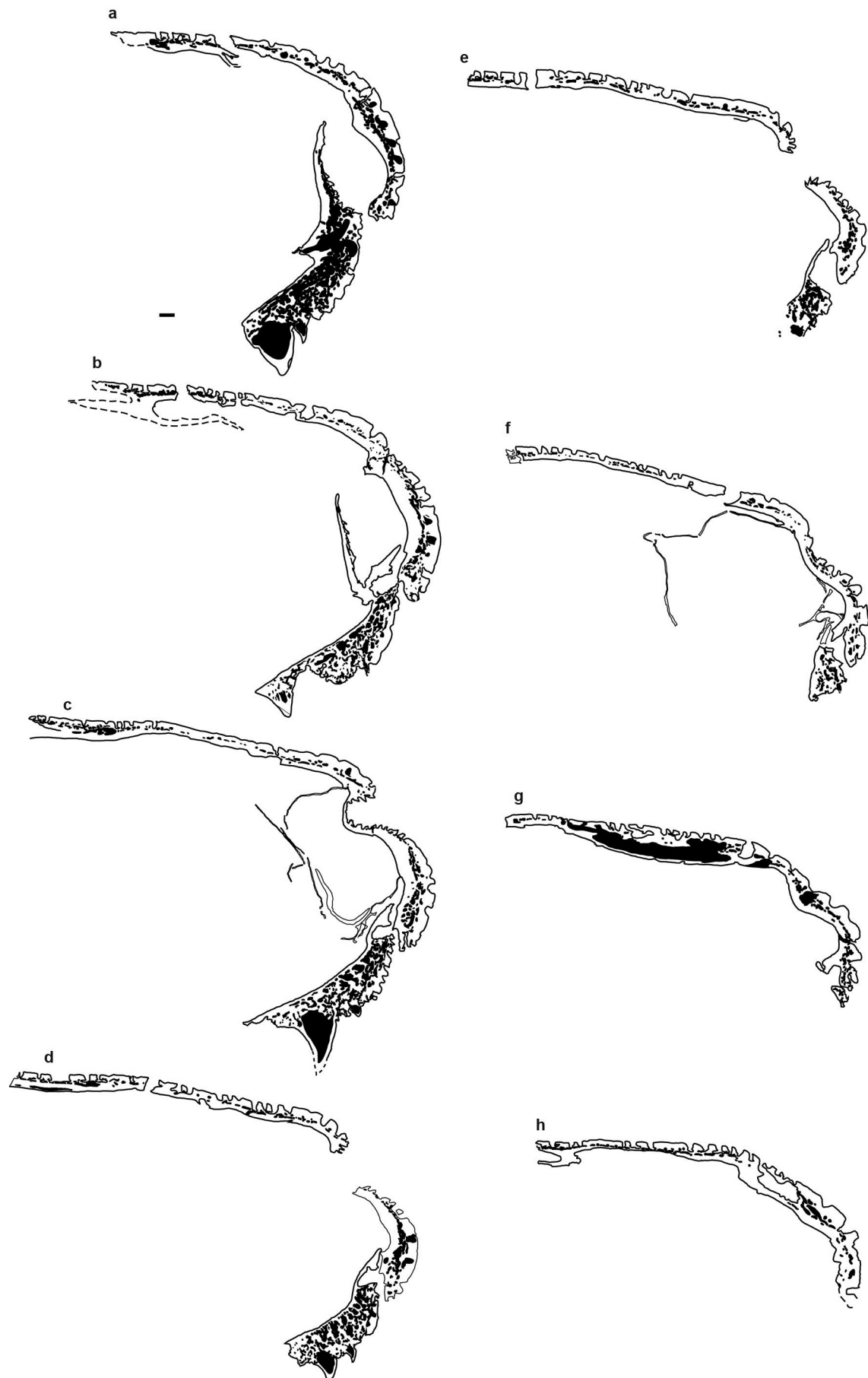
Extended Data Figure 1 | Exon organization of P/Q-rich SCPP genes in human, anole lizard, coelacanth, spotted gar and zebrafish. Each box represents a single exon. 5' and 3' untranslated regions are marked in white and protein-coding regions are marked in black. Location of the signal peptide (SP) is marked in yellow (EMPs), orange (maturation-stage proteins) and grey

(zebrafish SCPP6). P/Q-labelled exons contain at least 25% of Pro and Gln residues. Exons with aromatic residues Phe, Tyr and Trp are marked with Y/F. Conserved amino acid motifs are indicated on the top or inside the exon boxes. Nearly identical exons are marked by asterisk or circle.

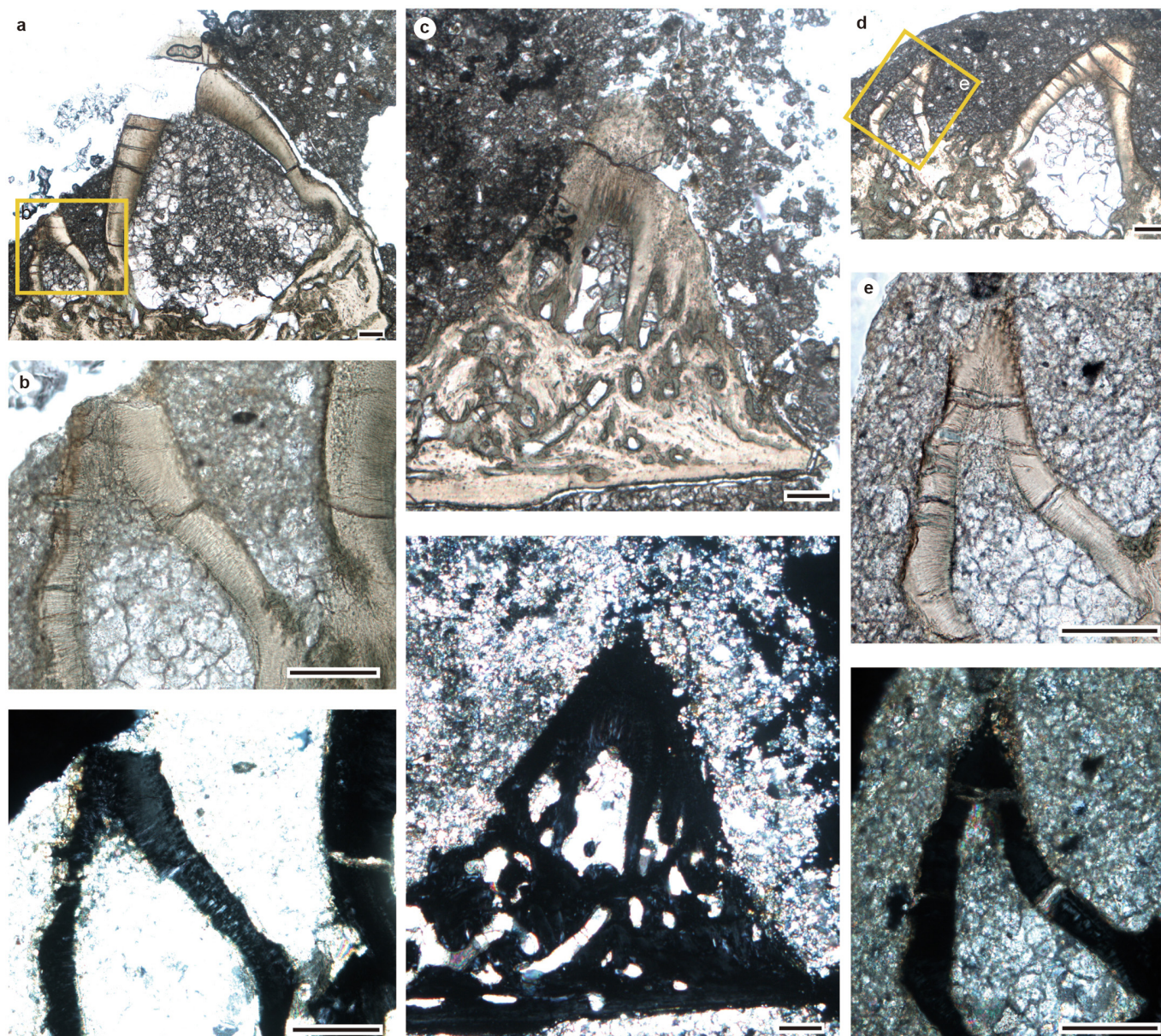


Extended Data Figure 2 | The skull roof (IVPP V17756) of *P. romeri* used for making thin sections in this study. Dorsal view (left) showing the relative positions of the thin sections and anterior view (right) showing the

position of premaxilla. 'a–h' represent positions of the eight sections in Extended Data Fig. 3. Scale bar, 5 mm.

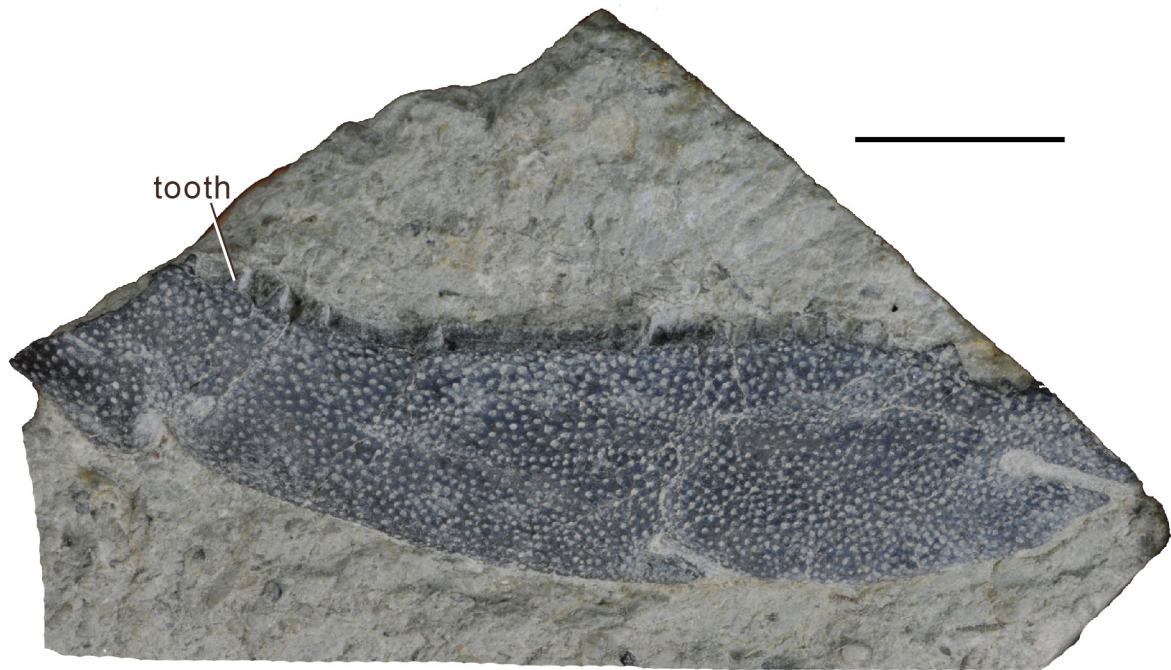


Extended Data Figure 3 | The outlines of all the thin sections made from IVPP V17756. Scale bar, 300 μm .



Extended Data Figure 4 | Other teeth on the left premaxilla showing the absence of enamel. **a**, From Extended Data Fig. 2a (IVPP V17756.1). **b**, Detail of **a** under transmission light (top) and polarized light (bottom). **c**, From Extended Data Fig. 2b (IVPP V17756.4), under transmission light (top) and

polarized light (bottom). **d**, From Extended Data Fig. 2d (IVPP V17756.4). **e**, Detail of **d** under transmission light (top) and polarized light (bottom). Scale bar, 100 μm.



Extended Data Figure 5 | The lower jaw (IVPP V19360) of *P. romeri* used for thin sections. Scale bar, 5 mm.

Extended Data Table 1 | P/Q-rich SCPP genes and SPARCL genes identified in the spotted gar and coelacanth genomes

| Gene | RNASeq transcript number in the skin | Number of paired-end RNASeq reads if ≥ 100 | Ensembl transcript number or GENSCAN | Signal peptide or predicted protein sequence |
|-----------|--------------------------------------|--|--------------------------------------|---|
| AMBN | RNASEQT00000075044 | skin ≈ 1700 | ENSLOCT00000016683 | MRSAILVLMCLIGTTLS |
| ENAM | RNASEQT00000075019 | skin ≈ 4900 | ENSLOCT00000016678 | MKIALLVLCFLGTSLA |
| AMTN | RNASEQT00000075051 | | | MKALIFMLCLVMSSG |
| SCPPPQ1 | RNASEQT00000075080 | skin ≈ 4700 | | MRTFILLACILPAVYL |
| SCPPPQ1* | | | GENSCAN prediction | MKAVFLLTCFLCAACAYPVPSSSMSSYSIPQFPPIPVQPFFPIPIQPVPPIPVQPQGPPIPVQP VPIPVQPQGPVPVPVQPPFLFPQPEAGIPFYGGYGGYGGGTGGGYGGYGGFPFRT |
| ODAM | RNASEQT00000075067 | skin ≈ 300 | ENSLOCT00000016696 | MHAALVFLSLVSTCLT |
| ODAM-like | RNASEQT00000075071 | skin ≈ 1700 | | MMKTVVLLVCFIGNIA |
| SCPP3A | RNASEQT00000075058 | | | MKTALFLMLLIGLSIA |
| SCPP3B | RNASEQT00000075064 | skin ≈ 400 | | MKTAIFLSCIHLTIA |
| SCPP3C | RNASEQT00000075062 | skin ≈ 200 | | MKTALFLMLLIGLSIA |
| SCPP5 | | skin ≈ 9900 larvae ≈ 700 embryo ≈ 6700 | GENSCAN prediction | MKTIILLTCFLFGSIFAAPMQPLYEFLPQVETPQQTAPYRQQPQQNNPYAPPSQTPQQQGPT RPASFEIMFPSQAFIRQKIPQAPGRQSIEILFPYSFGQHQQMFPYGNMPQTGGPQQNPFI NFGIPQVPVQEPTQPVQPVQPVQPVQPVQPSQNTPEVEEQD |
| SCPP7 | | skin ≈ 4000 | GENSCAN prediction | MKAVLLLAELIGSTLAIPMAQLYYDYNPALMAPQILQEPQQQIPDFVQGGEQSVGPARRAS FEILLPSRGFIKQSIQPGRPSLEILYPFGFGNTGPMGPVGPQLPALGTQTEPEKED |
| SCPP5/7 | RNASEQT00000075035 | skin ≈ 4600 embryo ≈ 1000 | | MQSVAVLVCLLGTTLA |
| SCPP9 | RNASEQT00000075075 | skin ≈ 100 | | MKTLTYFICFFQSFMHVLDA |
| SCPP10 | RNASEQT00000075078 | skin ≈ 700 | | MKSAILLYFFGAAFA |
| SPARCL1 | | eye ≈ 600 | ENSLOCT00000016704 | MMNLLLLCLLTAF |
| SPARCL2 | RNASEQT00000075010 | skin, liver, larvae, kidney, embryo ≈ 100 | | MKMLVLCLLLSPHLSS |

Ensembl annotated transcript numbers and RNA-seq transcript numbers together with GENSCAN predicted amino acid sequences for spotted gar and coelacanth used for analysis and construction of Fig. 2 and Extended Data Fig. 1. Predicted signal peptide sequences are marked in orange. First amino acid residue of each exon is highlighted in grey.

*Coelacanth.

Whole-genome sequencing identifies *EN1* as a determinant of bone density and fracture

A list of authors and affiliations appears at the end of the paper

The extent to which low-frequency (minor allele frequency (MAF) between 1–5%) and rare (MAF \leq 1%) variants contribute to complex traits and disease in the general population is mainly unknown. Bone mineral density (BMD) is highly heritable, a major predictor of osteoporotic fractures, and has been previously associated with common genetic variants^{1–8}, as well as rare, population-specific, coding variants⁹. Here we identify novel non-coding genetic variants with large effects on BMD ($n_{\text{total}} = 53,236$) and fracture ($n_{\text{total}} = 508,253$) in individuals of European ancestry from the general population. Associations for BMD were derived from whole-genome sequencing ($n = 2,882$ from UK10K (ref. 10); a population-based genome sequencing consortium), whole-exome sequencing ($n = 3,549$), deep imputation of genotyped samples using a combined UK10K/1000 Genomes reference panel ($n = 26,534$), and *de novo* replication genotyping ($n = 20,271$). We identified a low-frequency non-coding variant near a novel locus, *EN1*, with an effect size fourfold larger than the mean of previously reported common variants for lumbar spine BMD⁸ (rs11692564(T), MAF = 1.6%, replication effect size = +0.20 s.d., $P_{\text{meta}} = 2 \times 10^{-14}$), which was also associated with a decreased risk of fracture (odds ratio = 0.85; $P = 2 \times 10^{-11}$; $n_{\text{cases}} = 98,742$ and $n_{\text{controls}} = 409,511$). Using an *En1*^{cre/flox} mouse model, we observed that conditional loss of *En1* results in low bone mass, probably as a consequence of high bone turnover. We also identified a novel low-frequency non-coding variant with large effects on BMD near *WNT16* (rs148771817(T), MAF = 1.2%, replication effect size = +0.41 s.d., $P_{\text{meta}} = 1 \times 10^{-11}$). In general, there was an excess of association signals arising from deleterious coding and conserved non-coding variants. These findings provide evidence that low-frequency non-coding variants have large effects on BMD and fracture, thereby providing rationale for whole-genome sequencing and improved imputation reference panels to study the genetic architecture of complex traits and disease in the general population.

Recently, genetic discoveries have generally focused on common variants of small effect and rare coding variants identified through genome-wide association studies (GWAS) and whole-exome sequencing initiatives, respectively^{11,12}. The effect of low-frequency and rare non-coding variants upon common diseases, and their underlying traits has been recently explored in an isolated population^{13,14}, but has not been well-studied to date in the general population. The UK10K project has generated a large whole-genome sequence-based resource to address this question in a general European-ancestry population¹⁰, which is tenfold larger than the European subset of the 1000 Genomes project reference¹⁵.

Osteoporosis, diagnosed mainly through measurement of bone mineral density (BMD), is a common systemic skeletal disease characterized by an increased propensity to fracture. The narrow-sense heritability of BMD has been estimated to be $\sim 85\%$, and GWAS have successfully identified numerous loci associated with BMD which in total explain $\sim 5\%$ of the genetic variance for this trait¹⁶. However, these studies have been mainly unable to assess the role of low frequency (MAF 1–5%) and rare (MAF \leq 1%) genetic variation, as these

methods rely on testing common variants (MAF \geq 5%). A recent sequencing-based study identified a rare nonsense variant associated with BMD using 4,931 Icelandic subjects with low BMD and 69,034 population-based controls⁹. This coding variant, which disrupts the function of *LGR4*, appears to be confined to the Icelandic population.

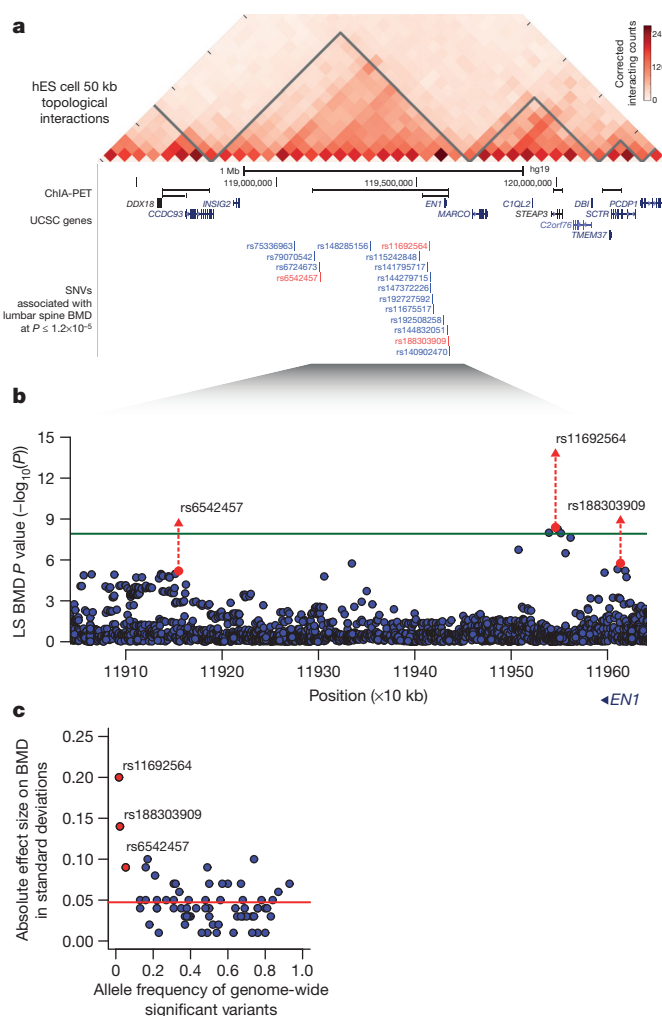


Figure 1 | Association signals near engrailed homeobox-1 for lumbar spine BMD. **a**, A topological domain includes associated variants and *EN1*, and chromatin interaction analysis with paired-end tag sequencing (ChIA-PET for CTCF in MCF-7 cell line) suggests a smaller interacting region containing *EN1*, and three genome-wide significant variants for lumbar spine BMD (in red). hES cell, human embryonic stem cell. **b**, Association signals at the *EN1* locus (green line at $P = 1.2 \times 10^{-8}$) for lumbar spine BMD. Red circles and triangles represent results from discovery and combined discovery and replication using fixed-effects meta-analysis (see Supplementary Information), respectively. **c**, Allele frequency versus absolute effect size for lumbar spine BMD for previously identified variants (blue)⁸ and the three *EN1* novel variants (red). The red line denotes the mean of previously reported effect sizes.

Table 1 | Novel variants from single SNV association tests

| BMD phenotype | SNP | Effect allele | Discovery meta-analysis | | | | Replication meta-analysis | | | | Combined meta-analysis | | | | |
|---------------|-------------|---------------|-------------------------|------|------------------------|----------------|---------------------------|------|-------------------------|----------------|------------------------|--------|------|-------------------------|----------------|
| | | | N | β | P | I ² | N | β | P | I ² | Freq. | N | β | P | I ² |
| Lumbar spine | rs11692564 | T | 25,225 | 0.24 | 4.1 × 10 ^{−9} | 0.37 | 15,291 | 0.20 | 2.8 × 10 ^{−6} | 0.46 | 0.016 | 40,516 | 0.22 | 1.7 × 10 ^{−14} | 0.40 |
| Lumbar spine | rs188303909 | T | 25,225 | 0.18 | 1.7 × 10 ^{−6} | 0.36 | 15,228 | 0.14 | 3.3 × 10 ^{−4} | 0.13 | 0.019 | 40,453 | 0.16 | 1.3 × 10 ^{−9} | 0.21 |
| Lumbar spine | rs6542457 | C | 25,225 | 0.08 | 6.5 × 10 ^{−6} | 0.00 | 15,240 | 0.09 | 1.5 × 10 ^{−4} | 0.00 | 0.058 | 40,465 | 0.09 | 2.2 × 10 ^{−9} | 0.00 |
| Femoral neck | rs55983207 | C | 29,188 | 0.10 | 2.5 × 10 ^{−7} | 0.19 | 16,248 | 0.17 | 9.8 × 10 ^{−10} | 0.03 | 0.042 | 45,436 | 0.12 | 7.2 × 10 ^{−15} | 0.23 |
| Femoral neck | rs11024028 | G | 29,188 | 0.06 | 2.2 × 10 ^{−8} | 0.00 | 15,397 | 0.03 | 2.6 × 10 ^{−2} | 0.30 | 0.198 | 44,585 | 0.05 | 1.3 × 10 ^{−9} | 0.04 |
| Forearm | rs148771817 | T | 7,848 | 0.47 | 9.3 × 10 ^{−9} | 0.15 | 2,539 | 0.41 | 5.5 × 10 ^{−4} | - | 0.012 | 10,387 | 0.46 | 1.1 × 10 ^{−11} | 0.00 |

β is the additive effect of the effect allele and is measured in standard deviations of bone mineral density.

To investigate the role of rare and low-frequency genetic variation on BMD in the general population of European descent, we first undertook whole-genome sequencing in 2,882 subjects from two cohorts in the UK10K project and whole-exome sequencing in 3,549 subjects from five cohorts (Supplementary Table 1) with BMD phenotypes. We then used a novel imputation reference panel generated by the UK10K and 1000 Genomes consortia to impute variants that were missing, or poorly captured, from previous GWAS studies in 26,534 subjects (Supplementary Table 1 and Extended Data Fig. 1a). The combined UK10K and 1000 Genomes reference panel, which contained 3,781 and 379 European individuals with whole-genome sequences from UK10K and 1000 Genomes projects, respectively, enabled improved imputation, particularly of low-frequency variants, when compared to the 1000 Genomes reference panel alone¹⁷. We then undertook *de novo* replication genotyping of lead variants in 13 cohorts for BMD, comprising 20,271 individuals of European descent.

We meta-analyzed association results from all discovery cohorts ($n_{\text{total}} = 32,965$, Supplementary Table 1) for BMD measured at the forearm, femoral neck and lumbar spine, the sites where osteoporotic fractures are most prevalent. We tested bi-allelic single nucleotide variants (SNVs) with $\text{MAF} \geq 0.5\%$ for association, declaring genome-wide statistical significance at $P \leq 1.2 \times 10^{-8}$ (accounting for all independent SNVs above this MAF threshold; Supplementary Methods)¹⁸. The sequence kernel association test (SKAT) was used to assess association of regions containing SNVs with $\text{MAF} \leq 5\%$ and $\leq 1\%$ (Supplementary Methods). All summary-level meta-analytic results are available for unrestricted download (<http://www.gefos.org>). Novel genome-wide significant loci were then tested for their relationship with fracture in up to 508,253 individuals. Finally, functional genomics as well as cellular and animal models were used to investigate the relevance of these novel genetic associations to bone physiology.

Through meta-analysis of sequenced and imputed single-SNV association tests from the discovery cohorts (Supplementary Table 1), we identified a novel locus at 2q14.2 harbouring variants associated with lumbar spine BMD (lead low-frequency SNV rs11692564(T), $\text{MAF} = 1.7\%$, effect size = $+0.24$ s.d., $P = 4 \times 10^{-9}$, Fig. 1 and Table 1). The direction of effect was consistent across all discovery cohorts (Extended Data Fig. 2) and the mean imputation information score for the imputed cohorts was 0.71 (Supplementary Table 2). This variant is located 53 kilobase pairs (kb) downstream from engrailed homeobox-1 (*EN1*), which, to our knowledge, has not previously been associated with any osteoporosis-related traits in humans. The rs11692564 variant was not present on HapMap imputation panels, nor on genotyping chips, underlining the importance of developing more comprehensive imputation reference panels.

To validate whole-genome sequencing genotypes at rs11692564, we genotyped 1,853 whole-genome sequenced subjects, and found all genotypes to be perfectly concordant (Supplementary Table 3). We validated imputation of rs11692564 in 3,601 imputed subjects through direct genotyping and observed that the association strengthened, and its statistical significance improved, as compared to imputed results (lumbar spine: imputed effect size = 0.22 s.d.; $P = 0.05$, genotyped effect size = 0.31 s.d.; $P = 0.004$) (Supplementary Table 4). We next sought additional evidence for the association at rs11692564 by

performing additional *de novo* genotyping in 16,233 independent individuals and found a similarly large effect size in this population (effect size = $+0.20$ s.d.; $P = 3 \times 10^{-6}$). Meta-analysis of the discovery and replication cohorts provided strong evidence for association ($P_{\text{combined-meta}} = 2 \times 10^{-14}$) (Table 1).

We also identified an additional association signal, arising from rs55983207 ($\text{MAF} = 4\%$), 17 kb downstream of rs11692564 ($r^2 = 0.001$) to be associated with femoral neck BMD from the combined meta-analysis ($P_{\text{meta}} = 7.2 \times 10^{-15}$, Table 1). A haplotype containing both effect alleles was not observed from within the UK10K whole-genome sequenced cohort (total number of haplotypes = 7,562).

In addition to rs11692564, we also observed two additional novel genome-wide significant variants for lumbar spine BMD near *EN1*, rs6542457 ($\text{MAF} = 5.8\%$) and rs188303909 ($\text{MAF} = 1.6\%$), which are 391 kb downstream and 67 kb upstream from rs11692564, respectively (Fig. 1b and Table 1). Variant rs188303909 was in moderate linkage disequilibrium (LD) with rs11692564 ($r^2 = 0.47$), and conditional analysis demonstrated that these two association signals were not independent (Supplementary Table 5). However, rs6542457 was in low LD with rs11692564 ($r^2 = 0.002$), and remained independent in conditional analyses (Supplementary Table 5). Overall, the *EN1* locus harbours multiple non-coding variants associated with lumbar spine

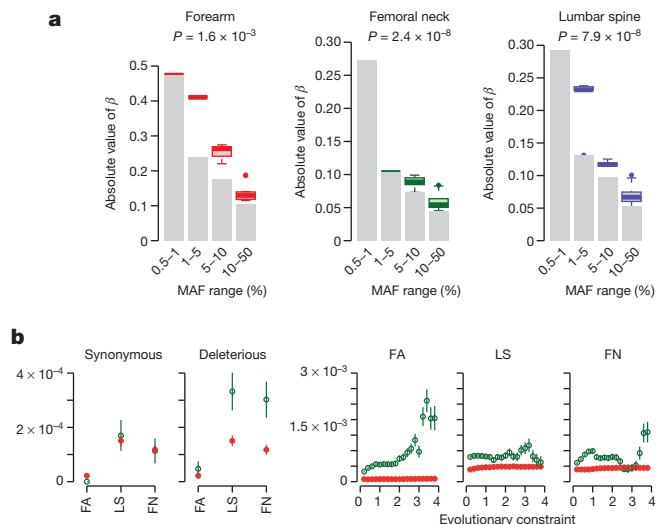


Figure 2 | Genome-wide features of association signals. **a**, Box plots of the effect sizes of genome-wide significant SNVs ($P < 1.2 \times 10^{-8}$), pruned for LD ($r^2 < 0.2$) by MAF bin for discovery cohorts. Grey bars represent the values of β not observed and for which we lack statistical power to observe (at $\alpha \leq 1.2 \times 10^{-8}$ and power ≥ 0.8). P values per phenotype are from the non-parametric trend test across MAF bins (see Supplementary Information). **b**, Proportion of single nucleotide variants (SNVs) passing a false discovery rate (FDR) q -value of 0.05 across different annotation features in discovery cohorts (green) versus matched control variants (red). The three panels on the right-hand side show enrichment across a range of evolutionary constraint scores (GERP++ score), in which green denotes SNVs above the threshold and red denotes variants below the threshold. Bars represent standard error (for Methods refer to the Supplementary Information). FA, forearm; FN, femoral neck; LS, lumbar spine.

Table 2 | Fracture meta-analysis of *EN1* variants

| Locus | SNP | Effect allele | Effect allele freq. | OR (95% CI) | P | N cases | N controls | I ² |
|------------|-------------|---------------|---------------------|------------------|-----------------------|---------|------------|----------------|
| <i>EN1</i> | rs11692564 | T | 0.02 | 0.85 (0.80–0.89) | 2.0×10^{-11} | 98,742 | 409,511 | 0.00 |
| | rs188303909 | T | 0.03 | 0.89 (0.85–0.93) | 9.8×10^{-7} | 95,669 | 405,697 | 0.00 |
| | rs55983207 | C | 0.05 | 0.93 (0.90–0.96) | 5.4×10^{-6} | 97,651 | 407,487 | 0.20 |
| | rs6542457 | C | 0.06 | 0.98 (0.95–1) | 1.2×10^{-1} | 95,669 | 405,697 | 0.17 |

and a single variant associated with femoral neck BMD. All three genome-wide significant variants for lumbar spine BMD (Table 1) co-localize solely with *EN1* in a sub-region of high interaction frequency within a single topologically associated domain¹⁹ (Fig. 1a).

The mean effect size of previously reported genome-wide significant single nucleotide polymorphisms (SNPs) (MAF $\geq 5\%$) from the largest GWAS meta-analysis to date for lumbar spine and femoral BMD was 0.048 s.d. and the largest effect size was 0.1 s.d.⁸. Hence, the observed effect size at rs11692564 is fourfold larger than this mean and twice that of the largest previously reported effect (Fig. 1c)⁸. For all

genome-wide significant variants, we observed larger effect sizes across decreasing MAF bins (Fig. 2a).

An increase in BMD is associated with a decrease in risk of bone fracture. We therefore tested the association of rs11692564(T) (the low-frequency allele at *EN1* associated with the largest increase in BMD) in 18 cohorts comprising 508,253 individuals (98,742 cases and 409,511 controls, Supplementary Table 6). rs11692564(T) was strongly associated with a decreased risk of fracture (odds ratio (OR) = 0.85 (95% confidence interval (CI): 0.80–0.89); $P = 2.0 \times 10^{-11}$; $I^2 = 0.00$) (Table 2 and Supplementary Table 7). Table 2 also shows clear associations between other variants near *EN1* and risk of fracture. The fracture association at rs11692564 was 2.9-fold larger than the mean of fracture associations detected in the largest GWAS to date, and 2.0-fold larger than the largest previously identified fracture association⁸.

EN1 encodes a homeobox gene central to mouse limb development²⁰, which has been shown to be involved in Wnt signalling interaction with Dkk1 (ref. 21). Studies of calvarial bone development and fracture healing of long bones in mice have shown that perinatal *En1*^{−/−} mutants display osteopenia and enhanced skull bone resorption²², whereas in normal adult mice *En1* is upregulated in the bone callus post-fracture²². Investigating the functional role of *EN1*, we detected *En1* expression during osteoblastogenesis in developing and mature cultured murine calvarial osteoblasts, but not in marrow-derived osteoclasts, or in human primary osteoclast cultures (Fig. 3a and Extended Data Fig. 3). To determine where *En1* is active in adult bones, we analysed vertebrae from *En1*^{LacZ/+} knock-in mice²³ and detected LacZ expression in proliferative and hypertrophic chondrocytes, osteogenic cells in the periosteum and trabecular bone surface, and in osteocytes of cortical and trabecular bone (Fig. 3b and Extended Data Fig. 4).

Using *En1*^{cre/+}; R26^{lox-STOP-lox-EYFP} reporter mice to genetically tag cells for which the *En1* promoter was active at any point within a cell

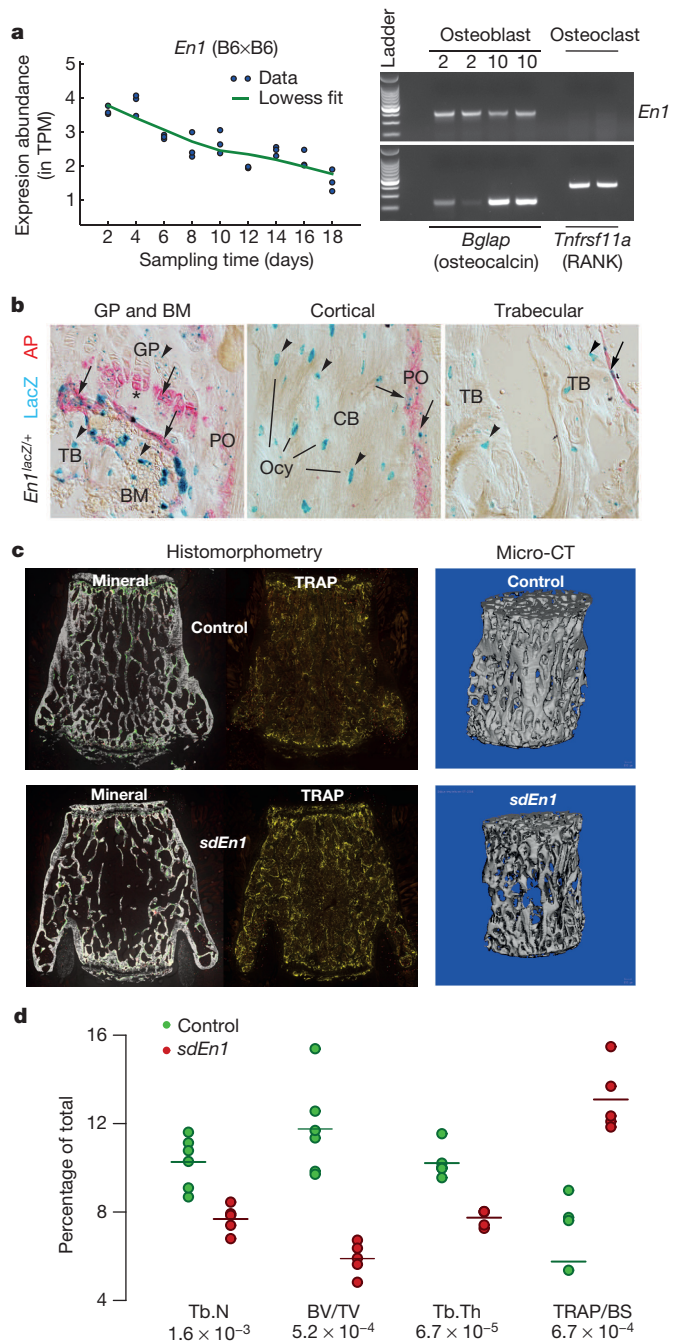


Figure 3 | Mouse *En1* functional experiments. **a**, Left, quantitative expression of *En1* and its temporal pattern (RNA-seq) in cultured calvarial murine osteoblasts ($n = 3$ per time point). Right, confirmation of the expression of *En1* in a separate RT-PCR experiment of cultured calvarial murine osteoblasts and lack of expression in osteoclasts matured from bone-marrow-derived precursor cells (positive controls for osteoblasts (osteocalcin) and osteoclast (RANK) are also shown). TPM, transcripts per million. **b**, Representative sections from lumbar vertebra 2 show the growth plate and bone marrow (GP and BM, left), cortical bone (CB, middle), and trabecular bone (TB, right) at $\times 40$ magnification from *En1*^{LacZ/+} adult mice ($n = 2$) stained for β -gal activity (LacZ blue, *En1*⁺ cells) and alkaline phosphatase (AP, red late chondrocytes and actively calcifying tissues). In the periosteum (PO), all the LacZ⁺ cells were AP⁺; some AP[−] BM cells expressed LacZ. Some AP⁺ proliferative chondrocytes in the GP expressed LacZ⁺, whereas most AP⁺ hypertrophic chondrocytes expressed LacZ. Some AP[−] osteocytes (Ocy) in CB and TB were LacZ⁺. **c**, Left, histomorphometry images of lumbar vertebrae 5 show decreased trabecular bone volume and increased bone surface area occupied by osteoclast cells when comparing *En1*^{cre/flox} (self-deleted *En1*, *sdEn1*) mutants and *En1*^{flox/+} control mice. Right, reconstructed micro-CT images show the mineral density in a control and an *sdEn1* animal. **d**, Micro-CT and histomorphometry measures within *sdEn1* ($n = 5$) and controls (*En1*^{flox/+}, $n = 6$). By micro-CT, *sdEn1* mutants exhibit decreased L5 trabecular number (Tb.N) and thickness (Tb.Th), as well as decreased bone volume fraction (BV/TV). Using histomorphometry, *sdEn1* mutants exhibit increased osteoclastic area (TRAP/BS). BS, bone surface; TRAP, tartrate acid staining. Average for each measure denoted by the solid horizontal line. For each group, P value between control and *sdEn1* is noted below label and was computed using paired t -test.

lineage, we confirmed that *En1* expression was only observed in osteogenic lineages (Extended Data Fig. 4). As most *En1*^{-/-} animals die soon after birth, we generated *En1*^{cre/flox} self-deleted *En1* (*sdEn1*) conditional mutants²⁴ ($n = 5$) and demonstrated by X-ray micro-computed tomography (micro-CT) that mutants have lower trabecular bone volume fraction (BV/TV), trabecular number, and trabecular thickness in both the lumbar L5 vertebrae (Fig. 3c, d and Extended Data Fig. 5) and the femur (Extended Data Fig. 5) as compared to littermate controls ($n = 6$). A decrease in femoral cortical thickness was also observed (Extended Data Fig. 5). By histomorphometry (Fig. 3c), we observed that the *sdEn1* mice had a statistically higher proportion of osteogenic and osteoclastic cells compared to littermate controls (Fig. 3d and Supplementary Table 8). The driving force for the low bone mass would appear to be an increase in osteoclastic activity induced by *En1* null osteogenic cells. This in turn initiates the expected coupled increase in mineralizing bone formation (Fig. 3b, d) mediated by an increased number of osteogenic cells and thus conforms to a high turnover osteoporosis-like phenotype, although dynamic histomorphometry and evidence from bone turn-over markers would be required to confirm an increased rate of bone formation (Extended Data Fig. 4). Genetic evidence from homologous regions in mice also supported a role for *En1* in bone, as the homologous region contained a quantitative trait loci (QTL) peak for femur BMD (Supplementary Table 9)²⁵. These findings, together with an earlier study focusing on *En1* function in calvarial bone development²² implicate this gene as an important mediator in skeletal biology.

Together, these findings suggest that *EN1* plays an important role in bone physiology and that low-frequency non-coding variants mapping near *EN1* have large effects on BMD and risk of fracture in the general European population.

We also identified a novel SNV at 7q31.31 within the intron of *CPED1* (rs148771817(T), MAF = 1.2%, effect size = +0.47 s.d., $P_{\text{discovery}} = 9.31 \times 10^{-9}$) associated with forearm BMD (Table 1, Supplementary Table 10 and Extended Data Fig. 6). We replicated the association at rs148771817 in 2,539 independent individuals and found a similar effect size (effect size = +0.41 s.d., $P = 6 \times 10^{-4}$), and combined meta-analysis of the discovery and replication cohorts for further improved statistical evidence for association (+0.46 s.d., $P = 1 \times 10^{-11}$) (Table 1). This variant had an effect size 2.2-fold larger than the mean of previously reported effects for common variants associated with forearm BMD (Extended Data Fig. 6)²⁶.

We previously identified rs7776725 to be associated with BMD at *WNT16*, a gene neighbouring *CPED1*, (Extended Data Fig. 6) and demonstrated that knockout of *Wnt16* in mice confers a 50% decrease in bone strength ($P = 7 \times 10^{-13}$)^{26,27}. We have recently shown that osteoblast-derived *Wnt16* represses osteoclastogenesis²⁸. As a result, we undertook conditional analysis of rs148771817 upon rs7776725. The rs148771817 variant remained associated after conditioning, albeit with lower statistical significance (effect size = 0.35 s.d.; $P_{\text{meta}} = 1 \times 10^{-7}$; Extended Data Fig. 6d). Similarly, conditional analysis of the common variant upon rs148771817 revealed little change in the effect size or the statistical significance (Supplementary Table 5). Although we acknowledge that both variants may be causal, our data does not permit us to distinguish if one or both of these variants have distinct biologic effects.

While rs148771817 is intronic in *CPED1*, we found that DNA accessibility at this region, as measured by DNase I hypersensitivity data from ENCODE studies, was moderately correlated with DNA accessibility at the *WNT16* promoter in 305 cell types (maximum $r^2 = 0.4$, $P = 2.2 \times 10^{-15}$, Supplementary Table 11), whereas correlation to the promoter of *CPED1* was lower (maximum $r^2 = 0.1$, $P = 0.06$). Moreover, analysis of chromosome conformation capture Hi-C interaction frequencies from human H1 embryonic stem cells shows elevated interaction frequency between rs148771817 and *WNT16* (Extended Data Fig. 6), though we also observed stronger interactions between these loci and their immediate neighbouring regions.

We assessed whether association signals were enriched for deleterious coding SNVs or SNVs with increased evolutionary constraint (see Supplementary Methods). These two groups of SNVs were matched to control SNVs by MAF and distance to gene (Supplementary Methods and Supplementary Table 12), followed by LD pruning ($r^2 < 0.2$). We observed enrichment of association signal across the spectrum of positive evolutionary constraint thresholds, which was comparable to deleterious coding variants (Fig. 2b).

In total, we have identified multiple variants associated with BMD, including 3 genome-wide significant loci for forearm BMD, 14 for femoral neck and 19 for lumbar spine (Supplementary Tables 10, 13–15, and Extended Data Figs 7 and 8). A common variant not on previous HapMap imputation panels, near the *SOX6* gene was also identified (rs11024028, MAF = 20%) (Table 1), and was found to be an independent signal from a previously reported signal at this locus (rs7108738, $r^2 = 0.002$)⁸. Consistent with recent experiments^{29,30}, region-based collapsing methods did not identify any convincing novel associations that were not already identified as genome-wide significant through single SNV associations. This included collapsing variants below 1% and 5% MAF thresholds, including all variants, only variants with increased GERP++ scores or those from protein-coding regions (Supplementary Table 16 and Extended Data Figs 9 and 10).

We have identified low-frequency, non-coding genetic variants of large effect that are present in the general population and associate with BMD and fracture. These variants have effect sizes up to fourfold larger than the mean effect described for common variants associated with BMD and approximately threefold larger than those for fracture. Our study illustrates that larger reference panels, covering relevant ethnicities, will facilitate the discovery of low frequency and rare variants. This was enabled here by a large imputation reference panel (UK10K and 1000 Genomes) which offered tenfold more European samples than the 1000 Genomes reference panel available at the time of analysis (phase I version 3). Although we did not identify coding low-frequency or rare variants associated with BMD at a genome-wide significant level, we did observe that deleterious coding variants were enriched for association as a group. This suggests the existence of as yet undiscovered coding variants influencing BMD. Importantly, we have also generated new functional evidence for a central role of the homeobox protein engrailed-1 gene in regulation of BMD and identified *EN1* as a critical protein in bone biology. Our findings demonstrate the utility of whole-genome sequencing-based discovery and deep imputation to enable the identification of novel genetic associations. These discoveries provide an improved understanding of the pathophysiology of osteoporosis and suggest that more comprehensive sets of whole-genome sequenced individuals, covering relevant ethnicities, will enable accurate imputation and thus facilitate discovery of low frequency and rare variants influencing complex traits and common disease.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 April 2014; accepted 30 June 2015.

Published online 14 September 2015.

- Richards, J. B. *et al.* Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet* **371**, 1505–1512 (2008).
- Styrkarsdottir, U. *et al.* Multiple genetic loci for bone mineral density and fractures. *N. Engl. J. Med.* **358**, 2355–2365 (2008).
- Styrkarsdottir, U. *et al.* New sequence variants associated with bone mineral density. *Nature Genet.* **41**, 15–17 (2009).
- Rivadeneira, F. *et al.* Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nature Genet.* **41**, 1199–1206 (2009).
- Duncan, E. L. *et al.* Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. *PLoS Genet.* **7**, e1001372 (2011).
- Koller, D. L. *et al.* Genome-wide association study of bone mineral density in premenopausal European-American women and replication in African-American women. *J. Clin. Endocrinol. Metab.* **95**, 1802–1809 (2010).

7. Xiong, D.-H. *et al.* Genome-wide association and follow-up replication studies identified *ADAMTS18* and *TGFB3* as bone mass candidate genes in different ethnic groups. *Am. J. Hum. Genet.* **84**, 388–398 (2009).
8. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature Genet.* **44**, 491–501 (2012).
9. Styrkarsdottir, U. *et al.* Nonsense mutation in the *LGR4* gene is associated with several human diseases and other traits. *Nature* **497**, 517–520 (2013).
10. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* <http://dx.doi.org/10.1038/nature14962> (this issue).
11. Hindorf, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
12. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nature Genet.* **44**, 623–630 (2012).
13. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genet.* **47**, 435–452 (2015).
14. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nature Genet.* **47**, 448–444 (2015).
15. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
16. Richards, J. B., Zheng, H.-F. & Spector, T. D. Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nature Rev. Genet.* **13**, 576–588 (2012).
17. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Comm.* **6**, 8111 (2015).
18. Xu, C., Tachmazidou, I., Walter, K., Ciampi, A., Zeggini, E. & Greenwood, C. M. T. Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* **38**, 281–290 (2014).
19. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
20. Loomis, C. A. *et al.* The mouse *Engrailed-1* gene and ventral limb patterning. *Nature* **382**, 360–363 (1996).
21. Adamska, M., MacDonald, B. T., Sarmast, Z. H., Oliver, E. R. & Meisler, M. H. *En1* and *Wnt7a* interact with *Dkk1* during limb development in the mouse. *Dev. Biol.* **272**, 134–144 (2004).
22. Deckelbaum, R. A., Majithia, A., Booker, T., Henderson, J. E. & Loomis, C. A. The homeoprotein engrailed 1 has pleiotropic functions in calvarial intramembranous bone formation and remodeling. *Development* **133**, 63–74 (2006).
23. Matisse, M. P. & Joyner, A. L. Expression patterns of developmental control genes in normal and *Engrailed-1* mutant mouse spinal cord reveal early diversity in developing interneurons. *J. Neurosci.* **17**, 7805–7816 (1997).
24. Sgaier, S. K. *et al.* Genetic subdivision of the tectum and cerebellum into functionally related regions based on differential sensitivity to engrailed proteins. *Development* **134**, 2325–2335 (2007).
25. Ackert-Bicknell, C. L. *et al.* Mouse BMD quantitative trait loci show improved concordance with human genome-wide association loci when recalculated on a new, common mouse genetic map. *J. Bone Miner. Res.* **25**, 1808–1820 (2010).
26. Zheng, H.-F. *et al.* *WNT16* influences bone mineral density, cortical bone thickness, bone strength, and osteoporotic fracture risk. *PLoS Genet.* **8**, e1002745 (2012).
27. Medina-Gomez, C. *et al.* Meta-analysis of genome-wide scans for total body BMD in children and adults reveals allelic heterogeneity and age-specific effects at the *WNT16* locus. *PLoS Genet.* **8**, e1002718 (2012).
28. Movérare-Skrtic, S. *et al.* Osteoblast-derived *WNT16* represses osteoclastogenesis and prevents cortical bone fragility fractures. *Nature Med.* **20**, 1279–1288 (2014).
29. Ladouceur, M., Dastani, Z., Aulchenko, Y. S., Greenwood, C. M. T. & Richards, J. B. The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. *PLoS Genet.* **8**, e1002496 (2012).
30. Tang, H. *et al.* A large-scale screen for coding variants predisposing to psoriasis. *Nature Genet.* **46**, 45–50 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements Full acknowledgements are listed in the Supplementary Information.

Author Contributions Principal Investigators: A.H., A.J., A.U., A.X.-A., B.L., C.A.-B., Ch.C., C.L., C.L.D., C.M.v.D., C.O., D.S.E., D.Ga., D.Go., D.Gr., D.H., D.Ki., D.M., E.D., E.O., F.Ri., F.Ro., G.D.S., J.B.R., J.D., J.Re., J.Ri., J.-T.K., J.Tu., K.A., L.A.C., L.L., L.P.G.M.d.G., M.B., M.M.F., N.S., N.T., N.v.d.V., N.v.S., P.R., R.D., R.L.P., S.G.W., S.H.R., T.H., T.P., T.S., U.P.-K., V.G., X.N., Y.-H.H. Genotyping: AOG Consortium, A.U., B.M., B.W., C.L., C.M.v.D., C.N., C.O., C.W., D.C., D.Gr., E.D., E.O., F.Ri., G.G., G.Tr., J.Er., J.J.v.M., J.Re., J.Ri., J.-T.K., J.v.R., M.B., M.C.F., M.J., M.Z., N.A., N.G.-G., N.S., N.T., P.Ar., P.D., P.R., R.K., S.H.R., S.M., S.R., U.P.-K., X.N. and Y.-H.H. Phenotyping: A.E., A.H., A.L., AOG Consortium, A.P.H., A.U., A.X.-A., B.M., C.G., C.K., C.L., C.L.D., C.M.v.D., C.O., D.Go., D.Ka., D.Ki., D.M., E.D., E.N., E.O., F.E.M., F.K., F.Ri., F.Ro., G.H., J.B., J.C., J.Ei., J.O., J.Re., J.Ri., J.-T.K., J.To., K.E., K.S., K.T., L.O., L.R., L.V., M.B., M.C.F., M.M.F., M.C.Z., M.Z., N.A., N.S., N.T., O.L., O.S., R.L.P., S.G.W., S.G., S.H.R., S.K., T.N., T.S. and U.P.-K. Functional experiments: A.J., A.R.-D., B.Ge., C.A.-B., C.H., C.L., C.L.D., C.O., C.U., D.Ga., D.P., E.G., H.Y.P.-M., J.D., J.F., K.Ch., Ma.M., M.H., N.S., O.S., S.B., S.C., S.-H.C., St.W., T.K., T.P., U.P.-K., W.C. and X.J. Data analysis: A.E., A.K., A.S., A.V.S., B.M., C.A.-B., Ch.C., C.-H.C., C.K., C.L., C.L.D., C.M.-G., C.M.T.G., C.O., C.T.L., C.W., D.S.E., D.M.E., D.C., D.Ka., D.M., D.P., E.D., E.G., E.N., F.G., F.Ri., G.H., G.Th., H.-F.Z., J.B.R., J.D., J.Er., J.F., J.Ha., J.Hu., J.K., J.v.R., K.Ch., K.E., K.W., L.A.C., L.H., L.M., L.O., L.R., L.V., M.B., M.C., M.H., M.K., N.A., N.S., N.T., O.L., P.Au., P.D., P.L., R.B., S.C., S.G.W., S.K., U.P., U.P.-K., V.F., W.-C.C., Y.-H.H., Y.M. and Y.Z. Meta-analysis: H.-F.Z., V.F. and Y.-H.H. Lead analysts: H.-F.Z. and V.F. Wrote first draft: J.B.R.

Author Information Source code used in preparation of results is available at <https://github.com/richardslab/gefos.seq/>. BMD discovery meta-analysis results are available from <http://www.gefos.org>. Information pertaining to UK10K can be obtained from <http://www.uk10k.org>. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.B.R. (brent.richards@mcgill.ca).

Hou-Feng Zheng^{1,2*}, Vincenzo Forgetta^{1,2*}, Yi-Hsiang Hsu^{3,4,5*}, Karol Estrada^{4,5,6,7*}, Alberto Rosello-Diez^{8*}, Paul J. Leo^{9*}, Chitra L. Dahia^{10,11*}, Kyung Hyun Park-Min^{12*}, Jonathan H. Tobias^{13,14*}, Charles Kooperberg^{15*}, Aaron Kleinman¹⁶, Unnur Styrkarsdottir¹⁷, Ching-Ti Liu¹⁸, Charlotta Uggla¹⁹, Daniel S. Evans²⁰, Carrie M. Nielson^{21,22}, Klaudia Walter²³, Ulrika Pettersson-Kymmer^{24,25}, Shane McCarthy²³, Joel Eriksson^{19,26}, Tony Kwan²⁷, Mila Jhamai⁶, Katerina Trajanoska^{6,28}, Yasin Memari²³, Josine Min¹⁴, Jie Huang²³, Petr Danecek²³, Beth Wilmot^{29,30}, Rui Li^{1,2}, Wen-Chi Chou^{3,4}, Lauren E. Mokry³¹, Alireza Moayyeri^{31,32}, Melina Claussnitzer^{3,4,5,33}, Chia-Ho Cheng³, Warren Cheung^{27,34}, Carolina Medina-Gómez^{6,28,35}, Bing Ge²⁷, Shu-Huang Chen²⁷, Kwangbom Choi³⁶, Ling Oei^{6,28,35}, James Fraser³⁷, Robert Kraaij^{6,28,35}, Matthew A. Hibbs^{36,38}, Celia L. Gregson³⁹, Denis Paquette³⁷, Albert Hofman^{28,35}, Carl Wibom⁴⁰, Gregory J. Tranah^{21,22}, Mhairi Marshall⁹, Brooke B. Gardiner⁹, Katie Cremin⁹, Paul Auer⁴¹, Li Hsu¹⁵, Sue Ring⁴², Joyce Y. Tung¹⁶, Gudmar Thorleifsson⁴³, Anke W. Ennenman⁶, Natasja M. van Schoor⁴⁴, Lisette C. P. G. M. de Groot⁴⁵, Nathalie van der Velde^{6,46}, Beatrice Melin⁴⁰, John P. Kemp^{9,14}, Claus Christiansen⁴⁷, Adrian Sayers³⁹, Yanhua Zhou¹⁸, Sophie Calderan^{48,49}, Jeroen van Rooij^{6,50}, Chris Carlson¹⁵, Ulrike Peters¹⁵, Soizik Berlivet⁵⁷, Joséé Dostie³⁷, Andre G. Uitterlinden^{6,28,35}, Stephen R. Williams⁵⁰, Charles Farber⁵⁰, Daniel Grinberg^{51,52,53}, Andrea Z. LaCroix⁵⁴, Jeff Haessler¹⁵, Daniel I. Chasman^{4,55}, Franco Giulianini⁵⁵, Lynda M. Rose⁵⁶, Paul M. Ridker^{4,55}, John A. Eisman^{56,57,58}, Tuan V. Nguyen^{56,58}, Jacqueline R. Center^{56,58}, Xavier Nogues^{59,60,61}, Natalia Garcia-Giralt^{59,60}, Lenore L. Launer⁶², Vilmundur Gudnason^{63,64}, Dan Mellström¹⁹, Liesbeth Vandenput¹⁹, Najaf Amin⁶⁵, Cornelia M. van Duijn⁶⁵, Magnus K. Karlsson⁶⁶, Östen Ljunggren⁶⁷, Olle Svensson⁶⁸, Göran Hallmans²⁵, François Rousseau^{69,70}, Sylvie Giroux⁷⁰, Johanne Bussière⁷⁰, Pascal P. Arp⁶, Fjorda Koromani^{6,28}, Richard L. Prince^{71,72}, Joshua R. Lewis^{71,72}, Bente L. Langdahl⁷³, A. Pernille Hermann⁷⁴, Jens-Erik B. Jensen⁷⁵, Stephen Kaptoge³¹, Kay-Tee Khaw⁷⁶, Jonathan Reeve^{77,78}, Melissa M. Formosa⁷⁹, Angela Xuereb-Anastasi⁷⁹, Kristina Åkesson^{66,80}, Fiona E. McGuigan⁸⁰, Gaurav Garg⁸⁰, Jose M. Olmos^{81,82}, Maria T. Zarrabeitia⁸³, Jose A. Riancho^{81,82}, Stuart H. Ralston⁸⁴, Nerea Alonso⁸⁴, Xi Jiang⁸⁵, David Goltzman⁸⁶, Tomi Pastinen^{27,34}, Elin Grundberg^{27,34}, Dominique Gauguier^{48,49}, Eric S. Orwoll^{22,87}, David Karasik^{3,88}, George Davey-Smith¹⁴, AOG Consortium†, Albert V. Smith^{32,71,91}, Evangelina E. Ntzani^{92,93}, Matthew A. Brown⁹, Kari Stefansson^{64,94}, David A. Hinds¹⁶, Tim Spector³², L. Adrienne Cupples^{18,95}, Claes Ohlsson¹⁹, Celia M. T. Greenwood^{2,34,96,97}, UK10K Consortium†, Rebecca D. Jackson⁹⁸, David W. Rowe⁸⁵, Cynthia A. Loomis⁹⁹, David M. Evans^{91,14}, Cheryl L. Ackert-Bicknell³⁶, Alexandra L. Joyner⁸, Emma L. Duncan^{91,100}, Douglas P. Kiel^{3,4,5,33}, Fernando Rivadeneira^{6,28,35} & J. Brent Richards^{1,2,32}

¹Departments of Medicine, Human Genetics, Epidemiology and Biostatistics, McGill University, Montréal H3A 1A2, Canada. ²Department of Medicine, Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montréal H3T 1E2, Canada. ³Institute for Aging Research, Hebrew SeniorLife, Boston, Massachusetts 02131, USA. ⁴Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁵Broad Institute of MIT and Harvard, Boston, Massachusetts 02115, USA. ⁶Department of Internal Medicine, Erasmus Medical Center, Rotterdam 3015GE, The Netherlands. ⁷Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁸Developmental Biology Program, Sloan Kettering Institute, New York, New York 10065, USA. ⁹The University of Queensland Diamantina Institute, Translational Research Institute, Princess Alexandra Hospital, Brisbane 4102, Australia. ¹⁰Department of Cell and Developmental Biology, Weill Cornell Medical College, New York, New York 10065, USA. ¹¹Tissue Engineering, Regeneration and Repair Program, Hospital for Special Surgery, New York 10021, USA. ¹²Rheumatology Division, Hospital for Special Surgery New York, New York 10021, USA. ¹³School of Clinical Science, University of Bristol, Bristol BS10 5NB, UK. ¹⁴MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, UK. ¹⁵Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. ¹⁶Department of Research, 23andMe, Mountain View, California 94041, USA. ¹⁷Department of Population Genomics, deCODE Genetics, Reykjavik IS-101, Iceland. ¹⁸Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02118, USA. ¹⁹Centre for Bone and Arthritis Research, Department of Internal Medicine and Clinical Nutrition, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg S-413 45, Sweden. ²⁰California Pacific Medical Center Research Institute, San Francisco, California 94158, USA. ²¹Department of Public Health and Preventive Medicine, Oregon Health & Science University, Portland, Oregon 97239, USA. ²²Bone & Mineral Unit, Oregon Health & Science University, Portland, Oregon 97239, USA. ²³Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK. ²⁴Departments of Pharmacology and Clinical Neurosciences, Aarhus University, Umeå S-901 87, Sweden. ²⁵Department of Public Health and Clinical Medicine, Umeå University, Umeå SE-901 87, Sweden. ²⁶Centre for Bone and Arthritis Research, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg S-413 45, Sweden. ²⁷McGill University and Genome Quebec Innovation Centre, Montréal H3A 0G1, Canada. ²⁸Department of Epidemiology, Erasmus Medical Center, Rotterdam 3015GE, The Netherlands. ²⁹Oregon Clinical and Translational Research Institute, Oregon Health & Science University, Portland, Oregon

- 97239, USA.³⁰Department of Medical and Clinical Informatics, Oregon Health & Science University, Portland, Oregon 97239, USA.³¹Farr Institute of Health Informatics Research, University College London, London NW1 2DA, UK.³²Department of Twin Research and Genetic Epidemiology, King's College London, London SE1 7EH, UK.³³Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02115, USA.³⁴Department of Human Genetics, McGill University, Montréal H3A 1B1, Canada.³⁵Netherlands Genomics Initiative (NGI)-sponsored Netherlands Consortium for Healthy Aging (NCHA), Leiden 2300RC, The Netherlands.³⁶Center for Musculoskeletal Research, University of Rochester, Rochester, New York 14642, USA.³⁷Department of Biochemistry and Goodman Cancer Research Center, McGill University, Montréal H3G 1Y6, Canada.³⁸Department of Computer Science, Trinity University, San Antonio, Texas 78212, USA.³⁹Musculoskeletal Research Unit, University of Bristol, Bristol BS10 5NB, UK.⁴⁰Department of Radiation Sciences, Umeå University, Umeå S-901 87, Sweden.⁴¹School of Public Health, University of Wisconsin, Milwaukee, Wisconsin 53726, USA.⁴²School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK.⁴³Department of Statistics, deCODE Genetics, Reykjavik IS-101, Iceland.⁴⁴Department of Epidemiology and Biostatistics and the EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam 1007 MB, The Netherlands.⁴⁵Department of Human Nutrition, Wageningen University, Wageningen 6700 EV, The Netherlands.⁴⁶Department of Internal Medicine, Section Geriatrics, Academic Medical Center, Amsterdam 1105, The Netherlands.⁴⁷Nordic Bioscience, Herlev 2730, Denmark.⁴⁸Cordeliers Research Centre, INSERM UMRS 1138, Paris 75006, France.⁴⁹Institute of Cardiometabolism and Nutrition, University Pierre & Marie Curie, Paris 75013, France.⁵⁰Departments of Medicine (Cardiovascular Medicine), Centre for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22908, USA.⁵¹Department of Genetics, University of Barcelona, Barcelona 08028, Spain.⁵²U-720, Centre for Biomedical Network Research on Rare Diseases (CIBERER), Barcelona 28029, Spain.⁵³Department of Human Molecular Genetics, The Institute of Biomedicine of the University of Barcelona (IBUB), Barcelona 08028, Spain.⁵⁴Women's Health Center of Excellence Family Medicine and Public Health, University of California – San Diego, San Diego, California 92093, USA.⁵⁵Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02215, USA.⁵⁶Osteoporosis & Bone Biology Program, Garvan Institute of Medical Research, Sydney 2010, Australia.⁵⁷School of Medicine Sydney, University of Notre Dame Australia, Sydney 6959, Australia.⁵⁸St. Vincent's Hospital & Clinical School, NSW University, Sydney 2010, Australia.⁵⁹Musculoskeletal Research Group, Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona 08003, Spain.⁶⁰Cooperative Research Network on Aging and Fragility (RETICEF), Institute of Health Carlos III, 28029, Spain.⁶¹Department of Internal Medicine, Hospital del Mar, Universitat Autònoma de Barcelona, Barcelona 08193, Spain.⁶²Neuroepidemiology Section, National Institute on Aging, National Institutes of Health, Bethesda, Maryland 20892, USA.⁶³Icelandic Heart Association, Kopavogur IS-201, Iceland.⁶⁴Faculty of Medicine, University of Iceland, Reykjavik IS-101, Iceland.⁶⁵Genetic epidemiology unit, Department of Epidemiology, Erasmus MC, Rotterdam 3000CA, The Netherlands.⁶⁶Department of Orthopaedics, Skåne University Hospital Malmö 205 02, Sweden.⁶⁷Department of Medical Sciences, University of Uppsala, Uppsala 751 85, Sweden.⁶⁸Department of Surgical and Perioperative Sciences, Umeå University, Umeå 901 85, Sweden.⁶⁹Department of Molecular Biology, Medical Biochemistry and Pathology, Université Laval, Québec City G1V 0A6, Canada.⁷⁰Axe Santé des Populations et Pratiques Optimales en Santé, Centre de recherche du CHU de Québec, Québec City G1V 4G2, Canada.⁷¹Department of Endocrinology and Diabetes, Sir Charles Gairdner Hospital, Nedlands 6009, Australia.⁷²Department of Medicine, University of Western Australia, Perth 6009, Australia.⁷³Department of Endocrinology and Internal Medicine, Aarhus University Hospital, Aarhus C 8000, Denmark.⁷⁴Department of Endocrinology, Odense University Hospital, Odense C 5000, Denmark.⁷⁵Department of Endocrinology, Hvidovre University Hospital, Hvidovre 2650, Denmark.⁷⁶Clinical Gerontology Unit, University of Cambridge, Cambridge CB2 2QQ, UK.⁷⁷Medicine and Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK.⁷⁸Institute of Musculoskeletal Sciences, The Botnar Research Centre, University of Oxford, Oxford OX3 7LD, UK.⁷⁹Department of Applied Biomedical Science, Faculty of Health Sciences, University of Malta, Msida MSD 2080, Malta.⁸⁰Clinical and Molecular Osteoporosis Research Unit, Department of Clinical Sciences Malmö, Lund University, 205 02, Sweden.⁸¹Department of Medicine and Psychiatry, University of Cantabria, Santander 39011, Spain.⁸²Department of Internal Medicine, Hospital U.M. Valdecilla- IDIVAL, Santander 39008, Spain.⁸³Department of Legal Medicine, University of Cantabria, Santander 39011, Spain.⁸⁴Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, Western General Hospital, University of Edinburgh, Edinburgh EH4 2XU, UK.⁸⁵Department of Reconstructive Sciences, College of Dental Medicine, University of Connecticut Health Center, Farmington, Connecticut 06030, USA.⁸⁶Department of Medicine and Physiology, McGill University, Montréal H4A 3J1, Canada.⁸⁷Department of Medicine, Oregon Health & Science University, Portland, Oregon 97239, USA.⁸⁸Faculty of Medicine in the Galilee, Bar-Ilan University, Safed 13010, Israel.⁸⁹Laboratory of Epidemiology, National Institute on Aging, National Institutes of Health, Bethesda, Maryland 20892, USA.⁹⁰Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.⁹¹School of Medicine and Pharmacology, University of Western Australia, Crawley 6009, Australia.⁹²Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece.⁹³Department of Health Services, Policy and Practice, Brown University School of Public Health, Providence, Rhode Island 02903, USA.⁹⁴deCODE Genetics, Reykjavik IS-101, Iceland.⁹⁵Framingham Heart Study, Framingham, Massachusetts 01702, USA.⁹⁶Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal H3A 1A2, Canada.⁹⁷Department of Oncology, Gerald Bronfman Centre, McGill University, Montréal H2W 1S6, Canada.⁹⁸Department of Medicine, Division of Endocrinology, Diabetes and Metabolism, The Ohio State University, Columbus, Ohio 43210, USA.⁹⁹The Ronald O. Perleman Department of Dermatology and Department of Cell Biology, New York University School of Medicine, New York, New York 10016, USA.¹⁰⁰Department of Diabetes and Endocrinology, Royal Brisbane and Women's Hospital, Brisbane 4029, Australia.
- *These authors contributed equally to this work.
 ‡These authors jointly supervised this work.
 †Lists of participants and their affiliations appear in the Supplementary Information.

METHODS

More details for the Methods are in the Supplementary Information. All human studies were approved by their institutional ethics review committees, and all participants provided written informed consent.

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment, except the teams undertaking micro-CT and histomorphometry experiments were blinded to each other's results.

Whole-genome sequencing. ALSPAC and TwinsUK cohorts were sequenced at an average read depth of 6.7× through the UK10K program (<http://www.UK10K.org>) using the Illumina HiSeq platform, and aligned to the GRCh37 human reference using BWA³¹. SNV calls were completed using samtools/bcftools and VQSR and GATK were used to recall these calls.

Whole-exome sequencing. The AOGC, FHS, RS-I, ESP and ERF cohorts were whole-exome sequenced as described in the Supplementary Information.

Whole-genome genotyping. All remaining discovery cohorts were genome-wide genotyped and imputed to the UK10K/1000 Genomes reference panel, as described in the Supplementary Information.

Association testing for BMD. Single variants with a MAF > 0.5% were tested for an additive effect on lumbar spine, femoral neck and forearm BMD, adjusting for sex, age, age², weight and standardized to have a mean of zero and a standard deviation of one. Meta-analysis of cohort-level summary statistics was undertaken using GWAMA³². Conditional analyses for significant SNVs was performed using GCTA³³. Region-based collapsing tests were performed using skatMeta³⁴, an implementation of the SKAT method³⁵ that enables the meta-analysis of multiple cohorts. For each cohort, variants with MAF ≤ 5% or ≤ 1% were collected and meta-analysis using skatMeta was conducted for windows of 30 SNVs within each region, overlapping by 10 SNVs.

Replication genotyping. Lead SNVs were selected for replication genotyping, which was performed at LGC Genomics, Erasmus MC and deCODE Genetics using KASP genotyping. Association testing for replication genotyping was undertaken using the same additive model, using the same covariates for BMD, as above.

Fracture association testing. Fractures were defined as those occurring at any site, except fingers, toes and skull, after age 18. Both incident and prevalent fractures were included and were verified by either radiographic, casting, physician, or subject reporting. Fractures resulting from any type of trauma were considered. Covariates included in the additive model were age, age², sex, height, weight, oestrogen/menopause status (when available), ancestral genetic background and cohort-specific covariates (such as clinical centre). Association testing was done in two phases. The first involved all 1,482 genome-wide significant SNVs for BMD. In the second phase of fracture association testing, variants at *EN1* were assessed in 18 cohorts, comprising 98,467 cases and 409,736 controls. Meta-analysis of cohort-level summary statistics was performed using GWAMA³².

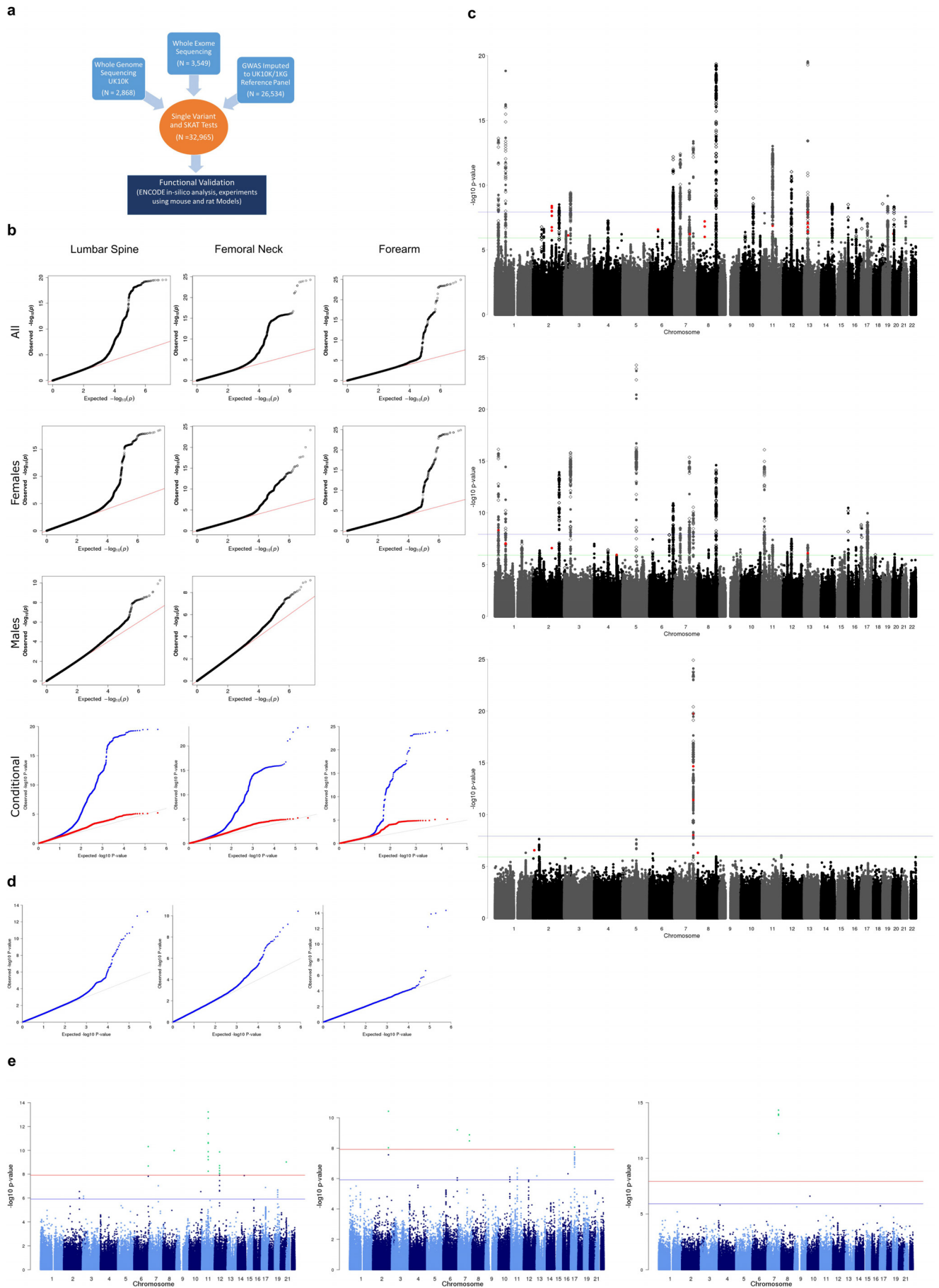
Functional genomics. We tested whether variants with increasing GERP++ scores³⁶ were more strongly associated with BMD than SNVs matched for distance to gene and MAF, after LD pruning using PLINK³⁷ at an r^2 of <0.2, using windows of 100 kb and a step of 20 kb. Coding variants were partitioned as deleterious using Variant Effect Predictor³⁸ LD pruned (r^2 < 0.2). The proportion of variants passing an FDR q -value of ≤ 0.05 were reported.

***En1* murine expression experiments.** Pre-osteoblast-like cell were differentiated to osteoblasts from calvaria of C57BL/6J mice and expression levels of each gene was quantified using RNA-seq. The temporal expression of *En1* in cell culture experiments of these osteoblasts and bone-marrow-derived osteoclasts (isolated from long bones of six-week-old mice) was measured by PCR, with *Bglap* (osteocalcin) and *Tnfrsf11a* (RANK), serving as controls. Total mRNA for *En1* in osteoblasts was quantified using real-time PCR.

Micro-CT and histomorphometry. Mouse husbandry and all experiments were performed in accordance with Memorial Sloan-Kettering Cancer Center Institutional Animal Care and Use Committee-approved protocols. Bone characteristics of self-deleted conditional *En1*(*sdEn1*) mutants were compared to *En1*^{+/flox} littermates using micro-CT. The same animals were assessed for histomorphometry (and laboratories performing micro-CT and histomorphometry were blinded to each other's results). After tissue sectioning, samples were stained for calcification (calcein blue), tartrate acid (TRAP) to assess for osteoclasts and alkaline phosphatase to assess for osteoblasts.

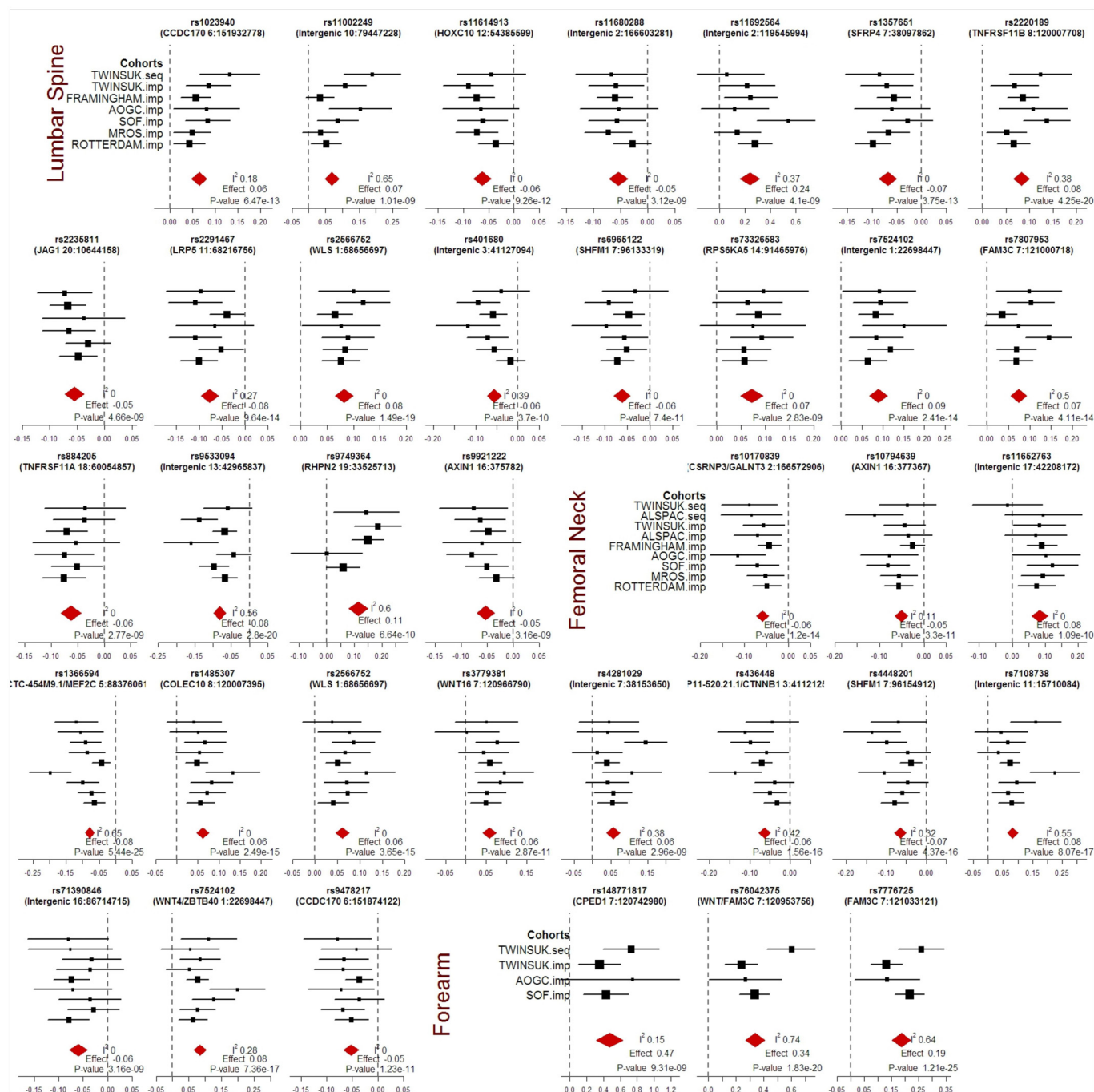
Murine histology. Two-month-old *En1*^{lacZ/+} mice³⁹ were sectioned at bone sites and stained for X-gal and/or alkaline phosphatase and imaged at ×400.

31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
32. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
33. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genet.* **44**, 369–375 (2012).
34. Voorman, A. A., Brody, J. & Lumley, T. SkatMeta: an R package for meta-analyzing region-based tests of rare DNA variants (<http://cran.r-project.org/web/packages/skatMeta>) (2013).
35. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
36. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput. Biol.* **6**, e1001025 (2010).
37. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
38. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics* **26**, 2069–2070 (2010).
39. Hanks, M., Wurst, W., Anson-Cartwright, L., Auerbach, A. B. & Joyner, A. L. Rescue of the *En-1* mutant phenotype by replacement of *En-1* with *En-2*. *Science* **269**, 679–682 (1995).



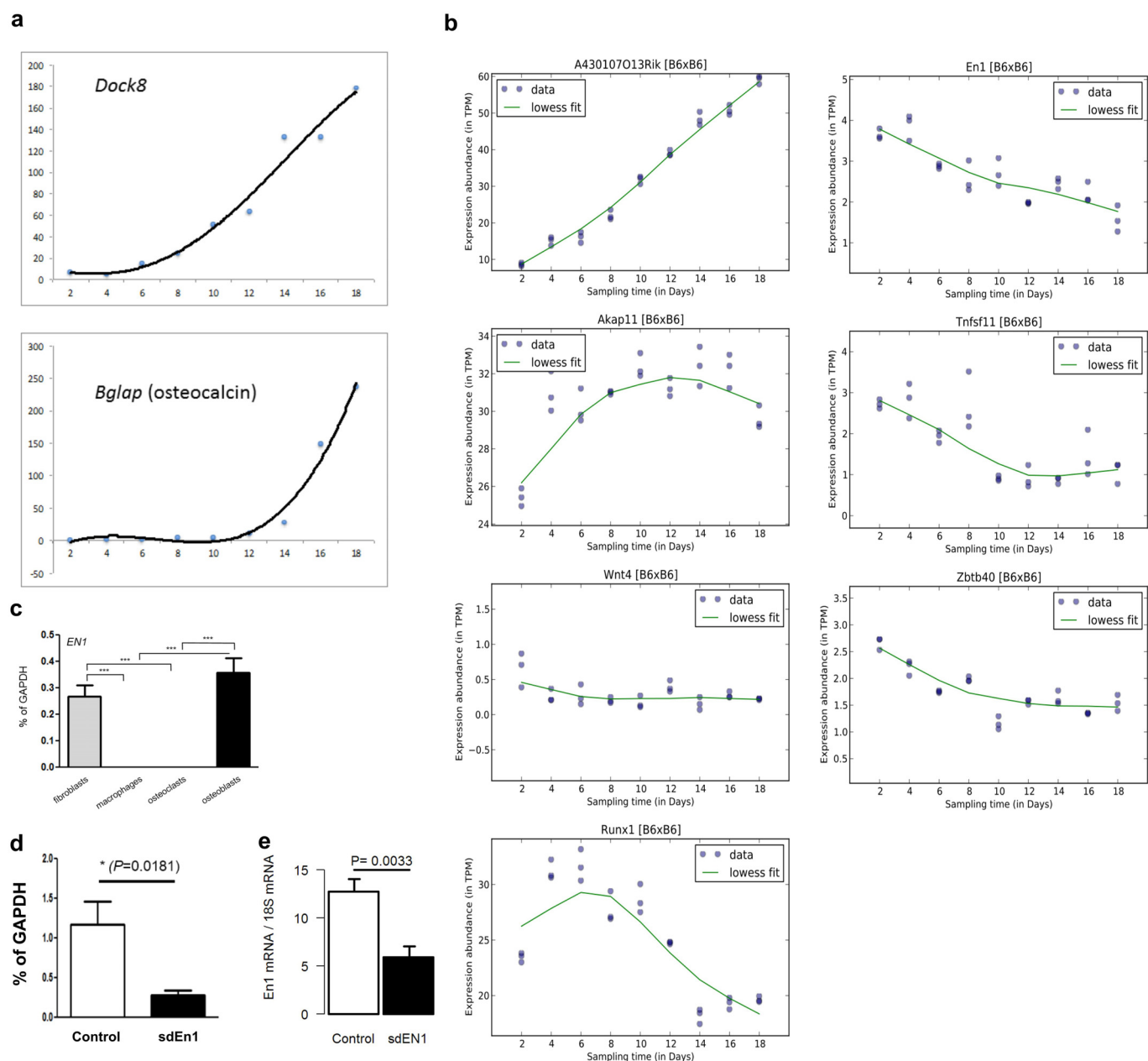
Extended Data Figure 1 | Discovery single variant meta-analysis. **a**, Overall study design. **b**, From top to bottom, quantile–quantile plots for the sex-combined single SNV meta-analysis, sex-stratified single SNV meta-analysis (forearm phenotype consists solely of female-only cohorts), and sex-combined single SNV conditional meta-analysis. Plots depict P values prior (blue) and after (red) conditional analysis on genome-wide significant variants (see Supplementary Methods). **c**, From top to bottom, Manhattan plots for sex-combined meta-analysis for lumbar spine BMD, femoral neck BMD, and forearm BMD. Each plot depicts variants from the UK10K/1000G reference panel with MAF $> 0.5\%$ across the 22 autosomes (odd, grey; even, black) against the $-\log_{10} P$ value from the meta-analysis of 7 cohorts (dots). Also

depicted are the subset variants from the reference panel that are also present in ref. 8 with P value $< 5 \times 10^{-6}$ (diamonds). Variants with MAF $< 5\%$ and $P < 1.2 \times 10^{-6}$ are also depicted (red). **d**, Quantile–quantile plots for the sex-combined meta-analysis of lumbar spine, femoral neck, and forearm BMD for SNVs present across both exome-sequenced and genome-sequenced and imputed cohorts, that is, SNV present only in genome-sequenced or imputed cohorts are not shown. **e**, Manhattan plot for the meta-analysis of sex-combined results for lumbar spine BMD for SNVs present in exome-sequenced and genome-sequenced and imputed cohorts, that is, SNV present only in genome-sequenced or imputed cohorts are not shown (from left to right: lumbar spine, forearm and femoral neck BMD).



Extended Data Figure 2 | Forest plots by cohort for genome-wide significant loci from discovery meta-analysis. Forest plots for three BMD phenotypes are shown. Title of each plot includes gene overlapping the SNP

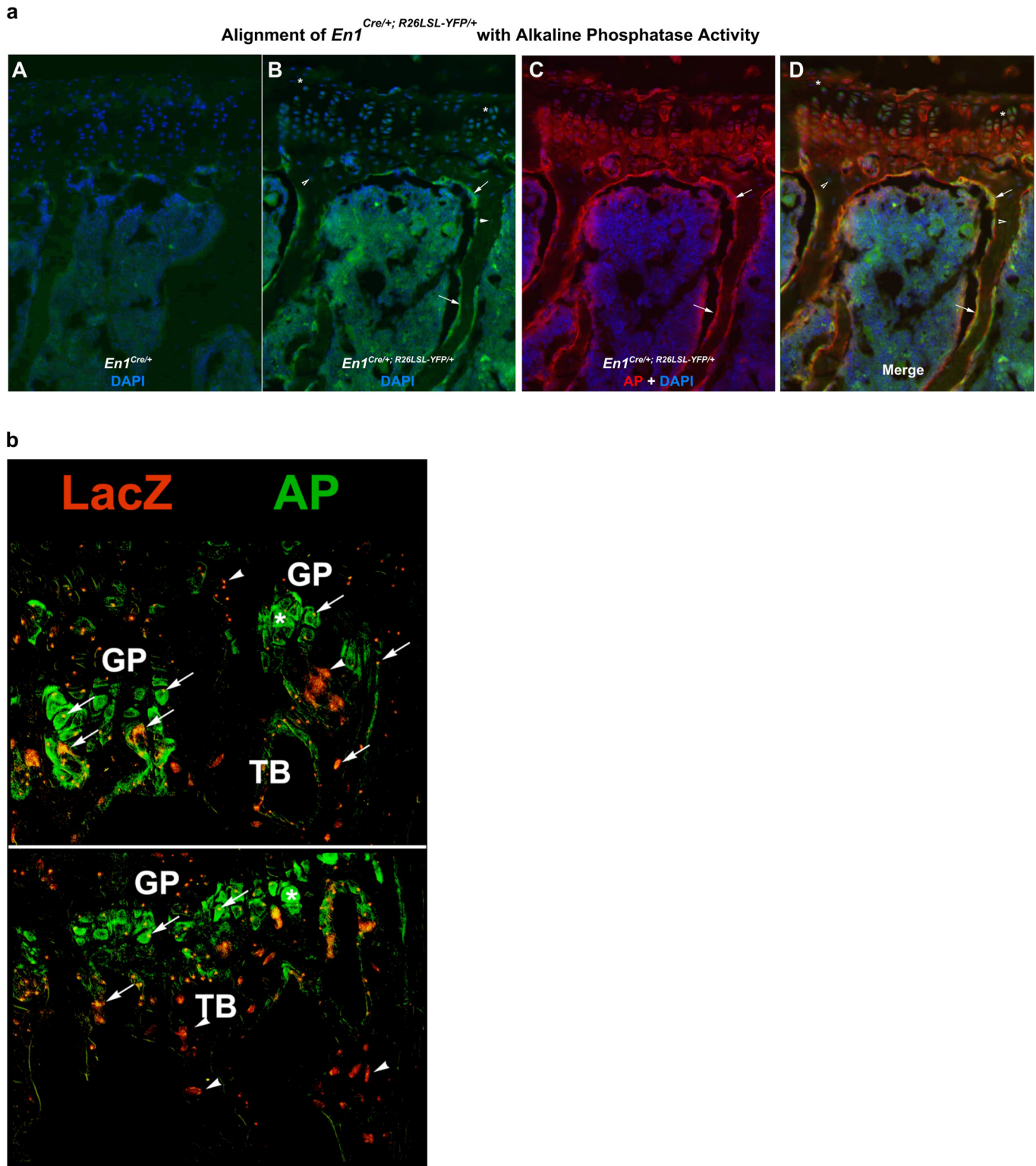
and its genomic position on build hg19. *P* values are from fixed-effect meta-analysis (see Supplementary Information).



Extended Data Figure 3 | Gene expression in human and mouse.

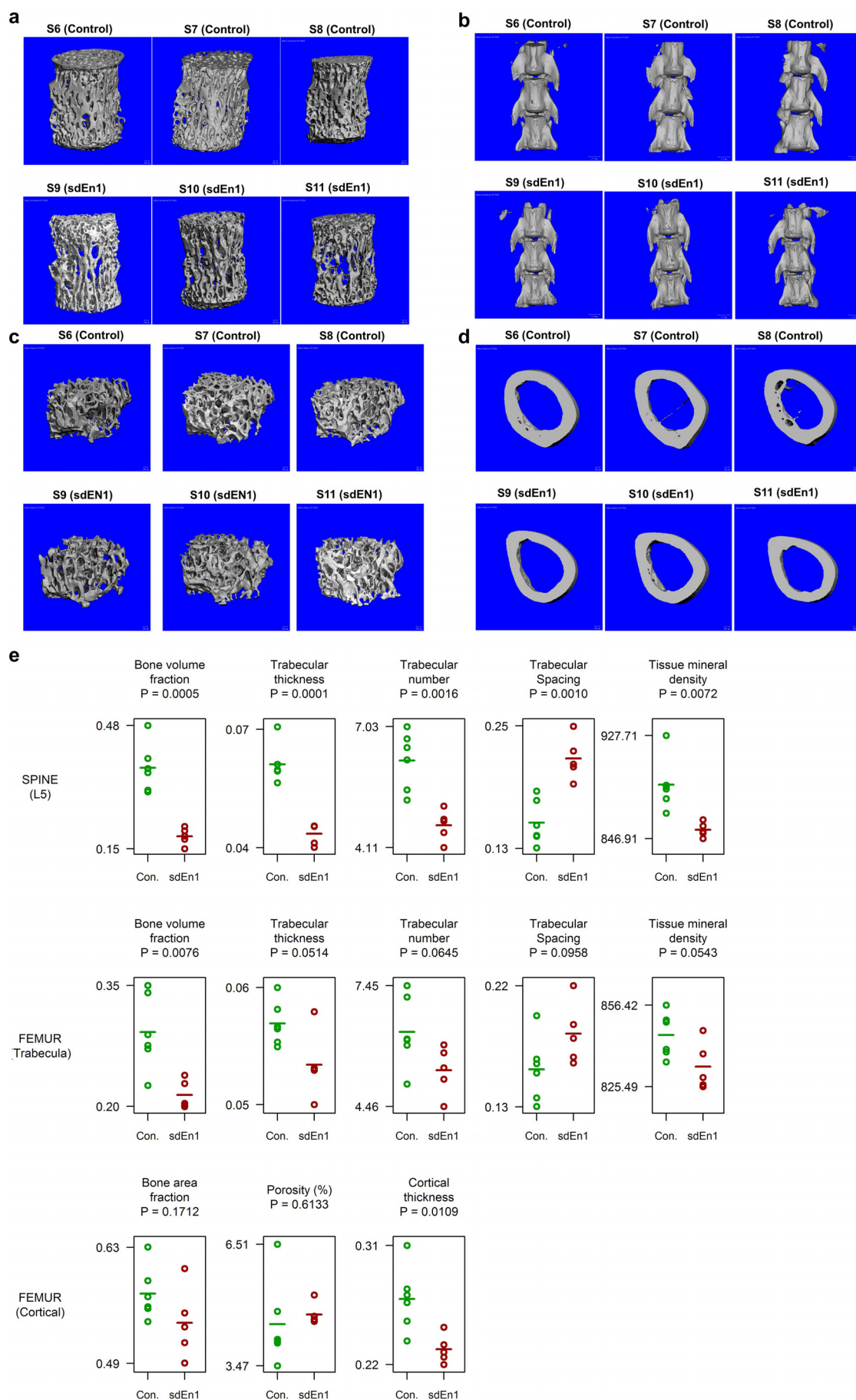
a, Quantification of *Dock8* expression and its temporal pattern through RNA-seq in cultured calvarial murine osteoblasts across day 2 through to day 18 of osteoblast development. Shown for comparison is *Bglap*, which encodes osteocalcin, a critical protein in osteoblasts. **b**, Quantification of expression of genome-wide significant genes and their temporal pattern through RNA-seq in cultured calvarial murine osteoblasts across day 2 through to day 18 of osteoblast development. **c**, Expression of *EN1* mRNA in human cells presented

as per cent of *GAPDH* mRNA. **d**, Expression of *En1* in control and *sdEn1* mice in purified osteoblast culture. For osteoblast marker gene expression, total mRNAs were purified from osteoblast cultures at day 10 and measured using quantitative real-time PCR. mRNA levels were normalized relative to *GAPDH* mRNA. **e**, Real-time PCR expression of control and *sdEn1* as compared to 18S mRNA in whole vertebral bone extract. All data are shown as mean \pm s.e.m. Significance computed by Student's unpaired *t*-test.

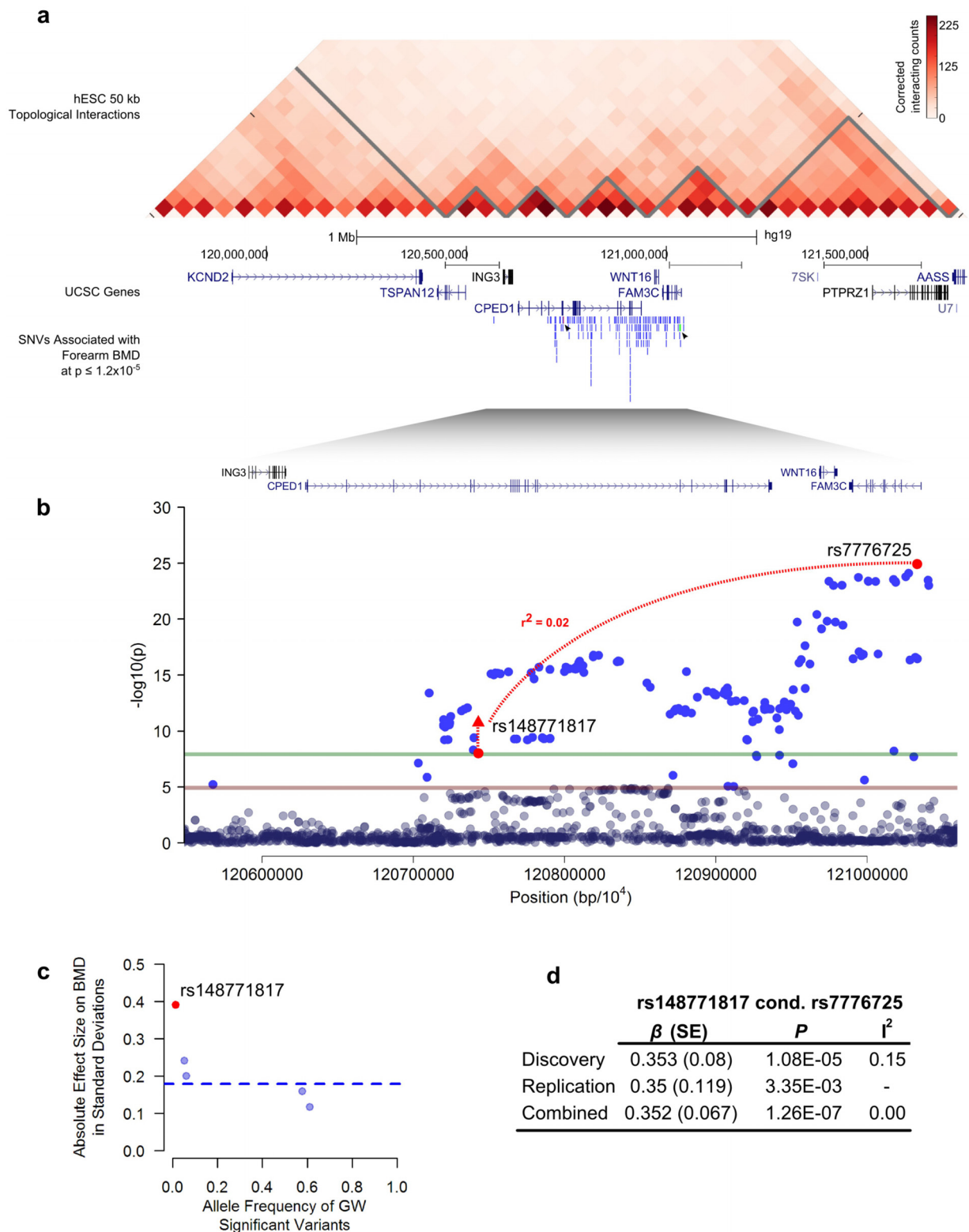


Extended Data Figure 4 | Histological assessment of *En1*^{Cre}-expressing cells in skeletal cells of the vertebra. **a**, Lineage history of *En1*^{Cre}-expressing cells in skeletal cells of the vertebra. The *En1*^{Cre} allele was combined with the *R26*^{LSL-YFP} reporter allele and examined using frozen fluorescent immunohistochemistry and alkaline phosphatase (AP) staining. Cell nuclei were detected with DAPI. YFP-expressing cells have expressed Cre (*En1*) at some time in their history. In subpanel A, control animals lacking the *R26*^{LSL-YFP} reporter show low background YFP signal (green). In subpanel B, *En1*^{Cre/+}; *R26*^{LSL-YFP/+} mice YFP-expressing cells are detected in the growth plate chondrocytes of the vertebra (asterisk), trabecular bone lining cells (arrow) and osteocytes (arrowhead). Note, high fluorescent background staining in the marrow space. In subpanel C, the same section is shown stained for AP activity using the Fast

Red substrate. Strong activity is present in the hypertrophic chondrocytes of the growth plate and trabecular bone lining cells (arrow). In subpanel D, alignment of the AP and YFP images shows that the trabecular lining cells co-express AP and YFP. **b**, Co-localization of *En1* and alkaline phosphatase expression. Images of lumbar vertebrae sections (growth plate and trabecular bone regions, $\times 40$ magnification) from two-month old *En1*^{lacZ/+} mice (see Fig. 3b), stained for LacZ and alkaline phosphatase (AP), false-coloured as indicated. Double-positive cells are indicated by arrows, single-positive cells are indicated by arrowheads (LacZ⁺) or asterisks (AP⁺). Except for some chondrocytes, most AP⁺ cells are also LacZ⁺, that is, express *En1*. The bone marrow was digitally removed, as it contains no AP⁺ cells.



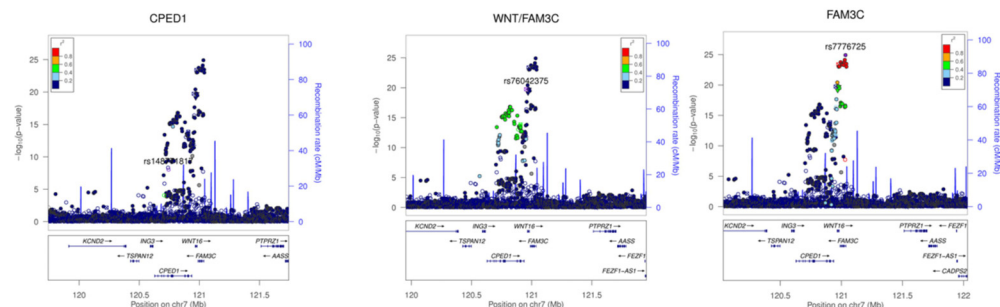
Extended Data Figure 5 | Micro-CT results for control (*En1^{flax/+}*) and self-deleting *En1* knockout (*sdEn1*, *En1^{cre/flax}*) animals. **a**, Trabecular bone micro-CT images from lumbar vertebra 5. **b**, Morphological characteristics at lumbar vertebra 4, 5, and 6 (from bottom to top). **c**, **d**, Morphological characteristics of left femur trabecular bone (**c**) and left femur cortical bone (**d**). **e**, Micro-CT parameter results for the comparison of control and *sdEn1* animals at lumbar vertebra 5, femur trabecula, and femur cortical bone. Horizontal lines denote mean of observations. Significance between control and *sdEn1* is calculated using an unpaired *t*-test.



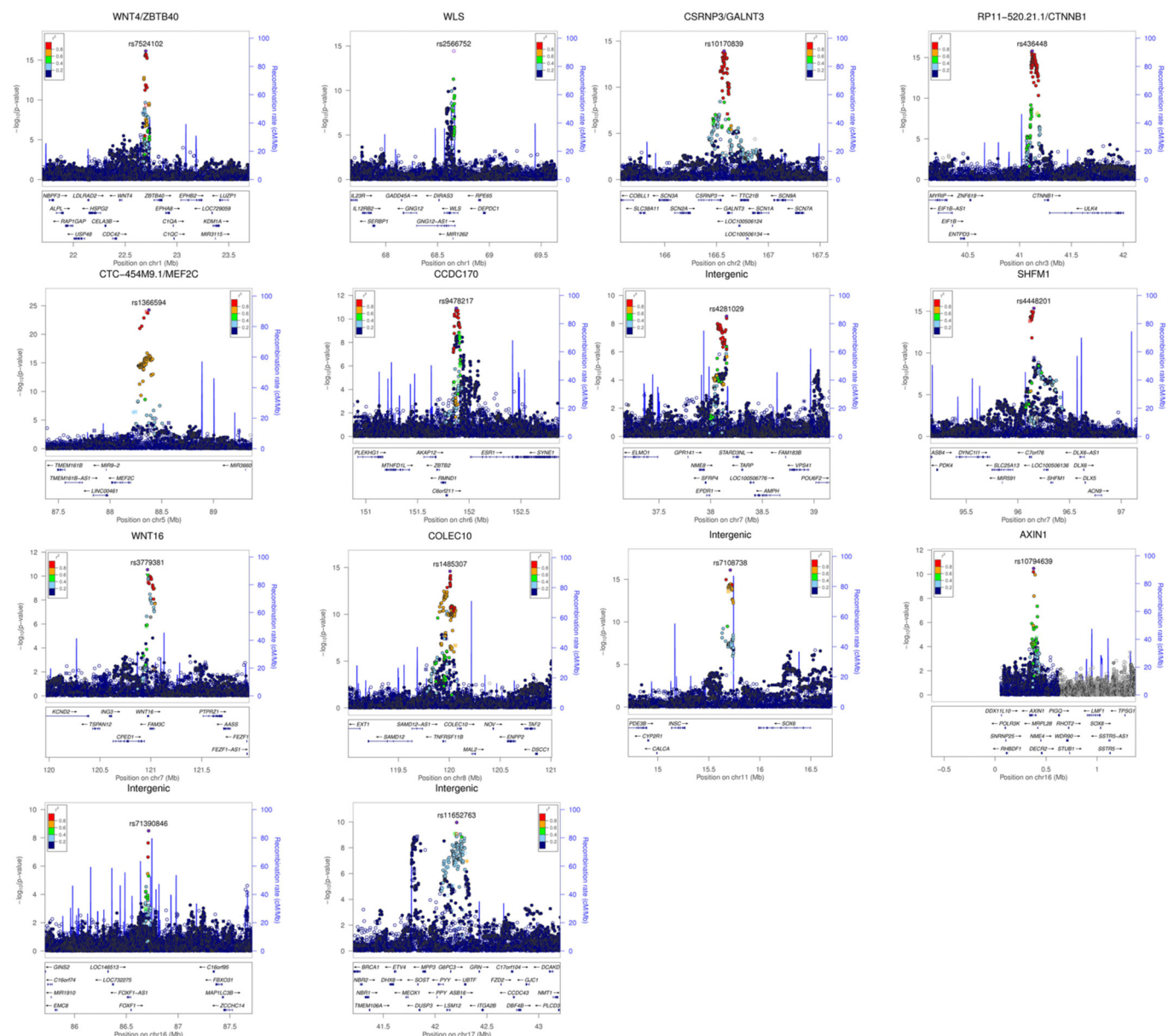
Extended Data Figure 6 | Novel association from 7q31.3. **a**, Chromatin interaction data from Hi-C performed in H1 embryonic stem cells²³ of a 2 Mb region encompassing rs148771817 (red and identified by arrow) and *WNT16*. **b**, The left axis denotes the association P value (red and green lines at $P = 1.2 \times 10^{-5}$ and 1.2×10^{-8} , respectively). The novel genome-wide significant SNV, rs148771817, within an intron of *CPED1*, and the lead genome-wide significant SNV rs7776725 upstream to *WNT16* (within

FAM3C) are in low LD with each other. **c**, Allele frequency versus absolute effect size (in standard deviations) for forearm BMD of all previously identified genome-wide significant variants (blue)⁸ and the novel variant within *CPED1* (red), rs148771817 from replication meta-analysis. The blue line denotes the mean of effect sizes for previously reported forearm BMD variants. **d**, Meta-analysis summary statistics of rs148771817 conditioned on rs7776725.

Forearm BMD



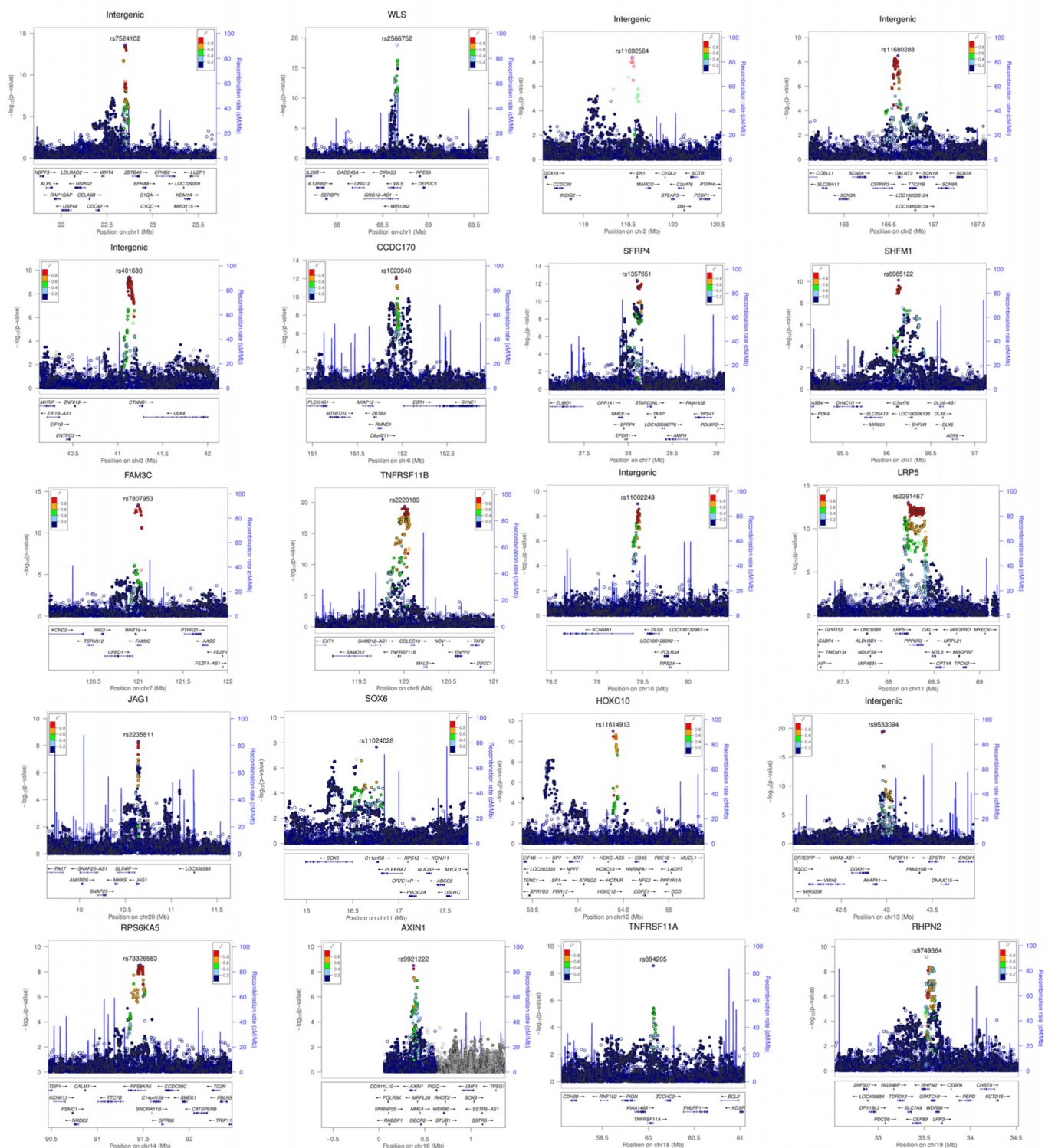
Femoral Neck BMD



Extended Data Figure 7 | Regional plots of genome-wide significant loci from single-SNV association tests for forearm and femoral neck BMD. Each regional plot depicts SNVs within 1 Mb of a locus' lead SNV (x axis) and their associated meta-analysis P value ($-\log_{10}$). SNVs are colour-coded

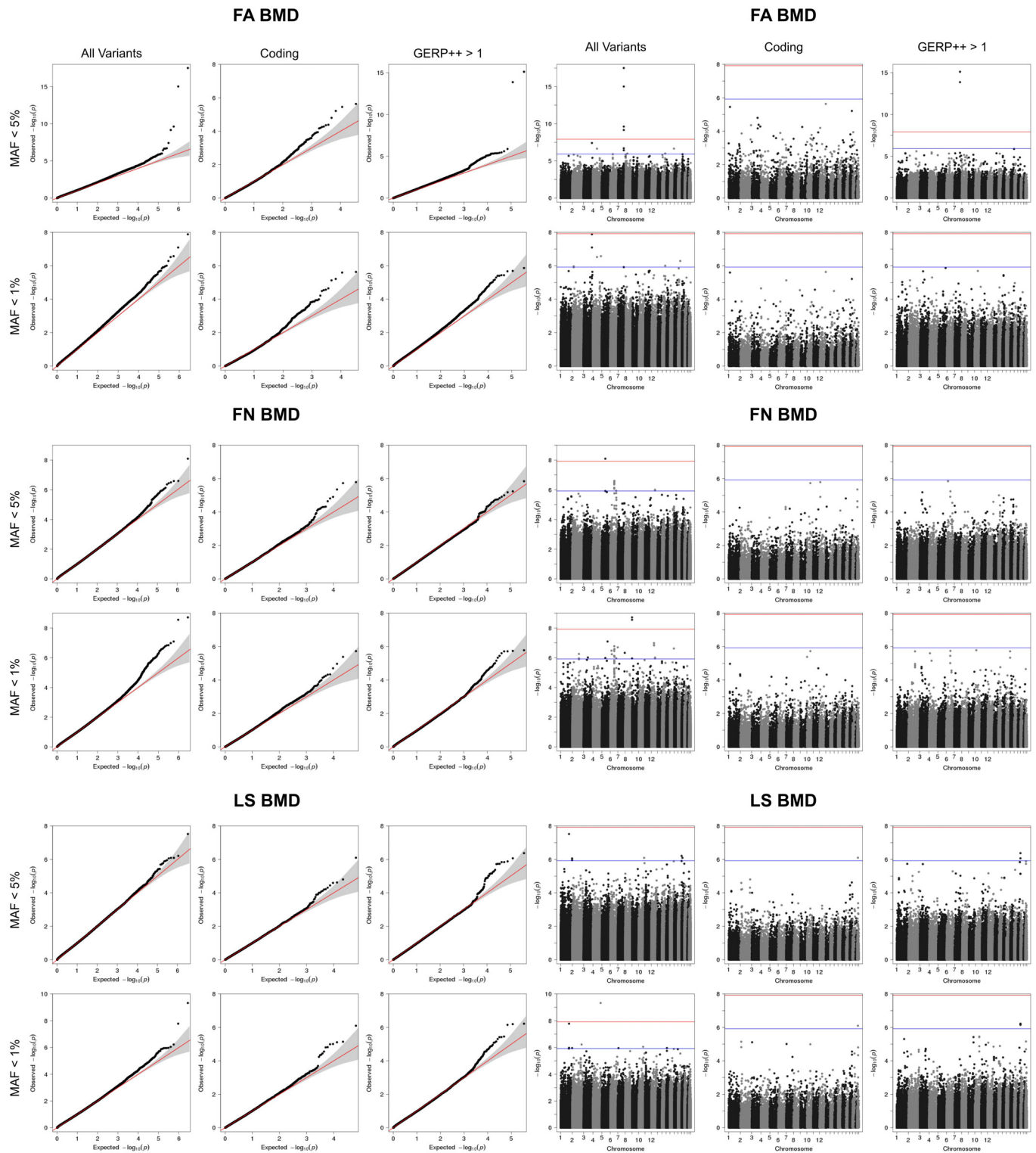
according to r^2 with the lead SNV (labelled, r^2 calculated from UK10K whole-genome sequencing data set). Recombination rate (blue line), and the position of genes, their exons and the direction of transcription are also displayed (below plot).

Lumbar Spine BMD



Extended Data Figure 8 | Regional plots of genome-wide significant loci from single-SNV association tests from lumbar spine BMD. Each regional plot depicts SNVs within 1 Mb of a locus' lead SNV (x axis) and their associated meta-analysis P value ($-\log_{10}$). SNVs are colour coded

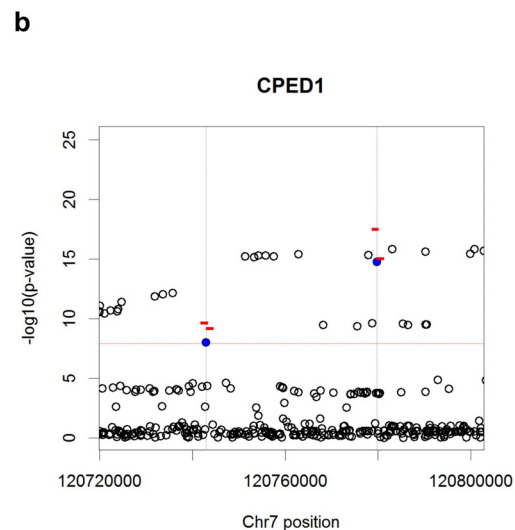
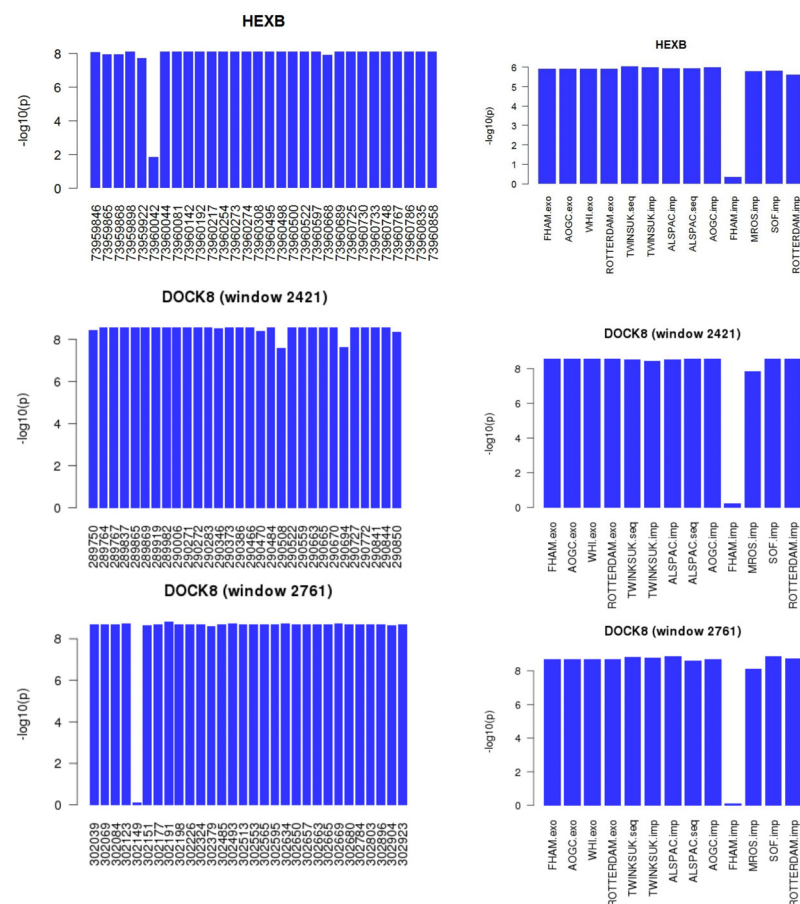
according to r^2 with the lead SNV (labelled, r^2 calculated from UK10K whole genome sequencing data set). Recombination rate (blue line), and the position of genes, their exons and the direction of transcription are also displayed (below plot).



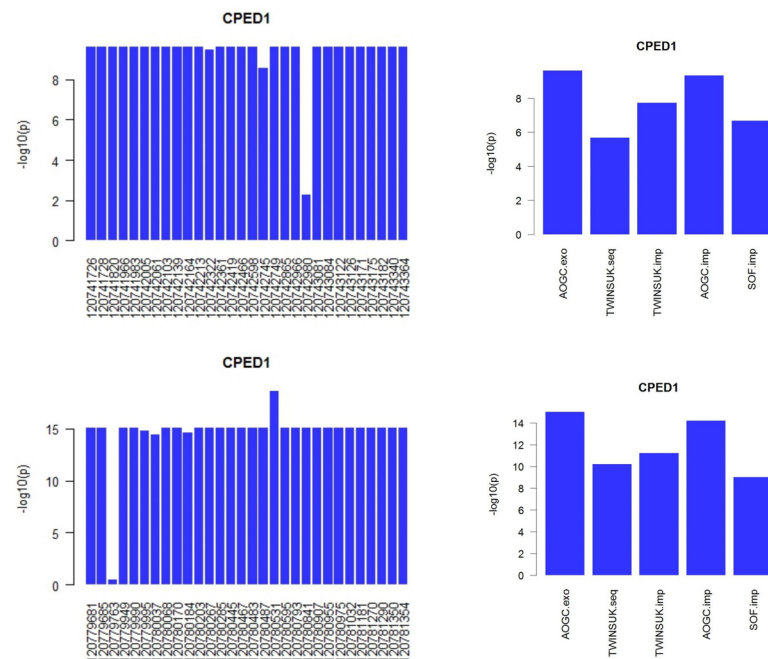
Extended Data Figure 9 | Region-based association tests using skatMeta for windows of 30 SNVs and window step of 20 SNVs. a, Left, quantile–quantile plots for forearm (FA) BMD, femoral neck (FN) BMD, and lumbar spine (LS) BMD. For each MAF range considered (<5% or <1%), analysis was conducted across all variants, variant overlapping coding exons, and variants with GERP++ score >1. b, Right, Manhattan plots forearm BMD, femoral neck

BMD, and lumbar spine BMD. For each MAF range considered (<5% or <1%), analysis was conducted across all variants, variant overlapping coding exons, and variants with GERP++ score >1. Blue lines indicate genome-wide suggestive ($P = 1.2 \times 10^{-6}$) thresholds and red lines indicate genome-wide significant ($P = 1.2 \times 10^{-8}$) thresholds.

a Femoral Neck BMD



Forearm BMD



Extended Data Figure 10 | Single variant analysis of signals from region-based tests. **a**, Drop-one SNV (left) and drop-one cohort (right) for genome-wide significant 30 SNV windows for femoral neck and forearm BMD from skatMeta analysis. On left, for a given 30 SNV window, the $-\log_{10}P$ of skatMeta test for 29 SNVs, excluding (that is, dropping) the SNV at position labelled on the x axis. On right, for given 30 SNV window on left,

the $-\log_{10}P$ of skatMeta test for all cohorts, excluding (that is, dropping) cohort labelled on x axis. **b**, Regional view of CPED1/WNT16 locus for forearm BMD. Significant SNVs from single variant meta-analysis (rs148771817 and rs79162867, in blue) overlap significant regions found using region-based test (red bars).

Nanoparticle biointerfacing by platelet membrane cloaking

Che-Ming J. Hu^{1,2*}, Ronnie H. Fang^{1,2*}, Kuei-Chun Wang^{3,4*}, Brian T. Luk^{2,3}, Soracha Thamphiwatana^{1,2}, Diana Dehaini^{1,2}, Phu Nguyen^{3,4}, Pavimol Angsantikul^{1,2}, Cindy H. Wen⁵, Ashley V. Kroll^{1,2}, Cody Carpenter¹, Manikantan Ramesh¹, Vivian Qu¹, Sherrina H. Patel⁵, Jie Zhu⁵, William Shi⁵, Florence M. Hofman⁶, Thomas C. Chen⁶, Weiwei Gao^{1,2}, Kang Zhang^{1,4,5,7}, Shu Chien^{3,4} & Liangfang Zhang^{1,2,4}

Development of functional nanoparticles can be encumbered by unanticipated material properties and biological events, which can affect nanoparticle effectiveness in complex, physiologically relevant systems^{1–3}. Despite the advances in bottom-up nanoengineering and surface chemistry, reductionist functionalization approaches remain inadequate in replicating the complex interfaces present in nature and cannot avoid exposure of foreign materials. Here we report on the preparation of polymeric nanoparticles enclosed in the plasma membrane of human platelets, which are a unique population of cellular fragments that adhere to a variety of disease-relevant substrates^{4–7}. The resulting nanoparticles possess a right-side-out unilamellar membrane coating functionalized with immunomodulatory and adhesion antigens associated with platelets. Compared to uncoated particles, the platelet membrane-cloaked nanoparticles have reduced cellular uptake by macrophage-like cells and lack particle-induced complement activation in autologous human plasma. The cloaked nanoparticles also display platelet-mimicking properties such as selective adhesion to damaged human and rodent vasculatures as well as enhanced binding to platelet-adhering pathogens. In an experimental rat model of coronary restenosis and a mouse model of systemic bacterial infection, docetaxel and vancomycin, respectively, show enhanced therapeutic efficacy when delivered by the platelet-mimetic nanoparticles. The multifaceted biointerfacing enabled by the platelet membrane cloaking method provides a new approach in developing functional nanoparticles for disease-targeted delivery.

Owing to their role as circulating sentinels for vascular damage and for invasive microorganisms, platelets have inspired the design of many functional nanocarriers^{8–13}. The multitude of platelet functions stem from a unique set of surface moieties responsible for immune evasion^{14,15}, subendothelium adhesion^{5,16}, and pathogen interactions^{6,7}. By adopting a cell membrane cloaking technique^{17–19}, we demonstrate the preparation of platelet membrane-cloaked nanoparticles (PNPs) consisting of a biodegradable polymeric nanoparticle core shielded entirely in the plasma membrane of human platelets. Several inherent platelet properties, including immunocompatibility, binding to injured vasculature and pathogen adhesion, as well as their therapeutic implications, were studied (Extended Data Fig. 1a).

PNPs were prepared by fusing human platelet membrane with 100-nm poly(lactic-co-glycolic acid) (PLGA) nanoparticles. Before platelet collection, blood and plasma samples were mixed with EDTA, which prevents platelet aggregation by deactivating fibrinogen-binding integrin α IIb β 3 (ref. 20). Platelets were then processed for nanoparticle membrane cloaking (Extended Data Fig. 1b). Physicochemical characterizations revealed that the final PNPs were

approximately 15 nm larger than the uncoated PLGA nanoparticles (bare NPs) and possessed an equivalent surface charge to that of platelets and platelet membrane-derived vesicles (platelet vesicles) (Fig. 1a). Transmission electron microscopy (TEM) visualization showed the formation of distinctive nanoparticulates and consistent unilamellar membrane coatings over the polymeric cores (Fig. 1b and Extended Data Fig. 2). Improved colloidal stability was observed with the PNPs compared to bare NPs (Fig. 1c), which is attributable to the stabilizing effect by the plasma membrane's hydrophilic surface glycans²¹. Translocation of platelet membrane protein content, including immunomodulatory proteins, CD47, CD55, and CD59^{14,15}, integrin components, α IIb, α 2, α 5, α 6, β 1, and β 3, and other transmembrane proteins, GPIb α , GPIV, GPV, GPVI, GPIX, and CLEC-2^{5,16}, onto the nanoparticles was examined by western blotting (Fig. 1d and Extended Data Fig. 3). Platelets derived from multiple protocols were prepared in parallel for comparison, and it was observed that the PNP preparation resulted in membrane protein retention and enrichment that was very similar across the different platelet sources (Extended Data Fig. 3). Notably, platelets derived from blood treated with heparin, an anticoagulant that inactivates thrombin rather than platelets, showed evidence of higher platelet activation including increased GPIb α cleavage and CLEC-2 oligomerization²². Using blood anticoagulated with EDTA as the platelet source, a right-side-out membrane orientation on the PNPs was verified by both immunogold staining and flow cytometric analysis with antibodies targeting either the intracellular or extracellular domain of CD47 (Fig. 1e and Extended Data Fig. 4). Pro-thrombotic, platelet-activating molecules such as thrombin, ADP and thromboxane were removed in the PNP formulation (Fig. 1f–h), thereby permitting PNP administration with little risk of a thrombotic response (Fig. 1i).

The platelet-mimicking functionalities of PNPs were first studied via binding of the particles to human type IV collagen, a primary subendothelial component²³. Fluorescently labelled PNPs, along with bare NPs and red blood cell membrane-cloaked nanoparticles (RBCNPs), were incubated on collagen-coated plates. The PNPs showed significantly enhanced retention compared to bare NPs and RBCNPs (Fig. 2a), indicating that the collagen adhesion was membrane-type-specific. Reduced PNP retention on non-collagen-coated plates and in the presence of anti-GPVI antibodies supports a specific collagen-platelet membrane interaction attributable to the presence of membrane glycoprotein receptors for collagen¹⁶ (Extended Data Fig. 3). Further examination of the differential binding of PNPs to endothelial and collagen surfaces was performed using collagen-coated tissue culture slides seeded with human umbilical vein endothelial cells (HUVECs). PNPs adhered primarily outside of areas encompassed by the cells (Fig. 2b and Extended Data Fig. 5a–g). In addition, the

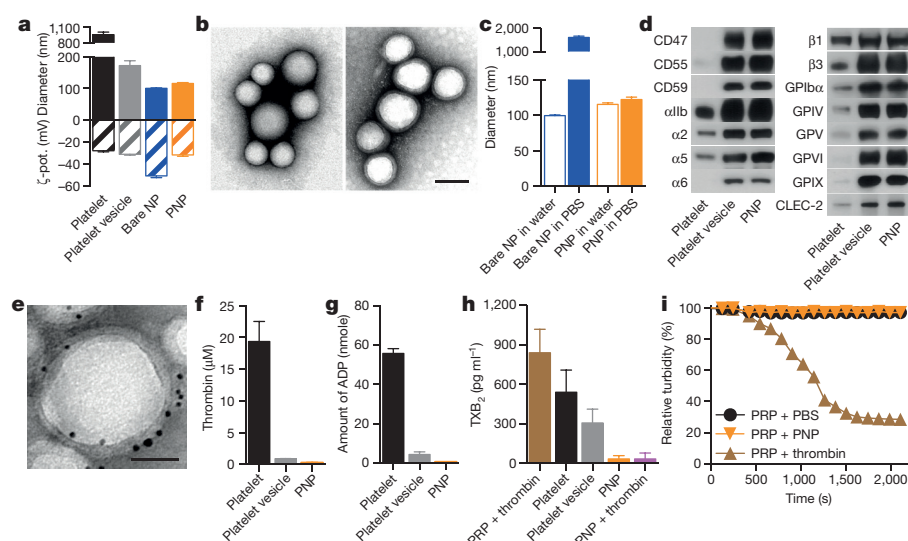
¹Department of NanoEngineering, University of California, San Diego, La Jolla, California 92093, USA. ²Moores Cancer Center, University of California, San Diego, La Jolla, California 92093, USA.

³Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, USA. ⁴Institute of Engineering in Medicine, University of California, San Diego, La Jolla, California 92093, USA.

⁵Shiley Eye Institute, University of California, San Diego, La Jolla, California 92093, USA. ⁶Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA.

⁷Veterans Administration Healthcare System, San Diego, California 92093, USA.

*These authors contributed equally to this work.



PNPs were incubated with the extracellular matrix derived from decellularized human umbilical cord arteries. After PBS washes, scanning electron microscopy (SEM) revealed a significant number of PNPs remaining on the fibrous structures on the luminal side of the artery (Fig. 2c and Extended Data Fig. 5h, i).

Examination of PNPs' immunocompatibility was conducted using differentiated human THP-1 macrophage-like cells. The platelet membrane cloaking reduced particle internalization in a CD47-specific manner²⁴, as blocking by anti-CD47 antibodies increased the cellular uptake (Fig. 2d). The PNPs were further investigated for their interactions with the complement system based on quantifications of C4d and Bb split products. After incubation in human plasma, complement activation was observed with bare NPs, reflecting their susceptibility to opsonization as well as the spontaneous reaction between C3 thioesters and the hydroxyl groups on the PLGA particles²⁵. In contrast, an equal amount

of PNPs mixed with autologous plasma showed no observable complement activation (Fig. 2e, f). This suppression of the complement system can be attributed to membrane-bound complement regulator proteins such as CD55 and CD59 (ref. 26; Extended Data Fig. 3). This result also attests to the completeness of the membrane cloaking, which shields the polymeric cores from plasma exposure and minimizes the risk of anaphylatoxin generation frequently associated with injectable nanocarriers²⁷.

The therapeutic potential of PNPs was first examined by assessing their selective adherence to damaged vasculatures. A segment of the human carotid artery was surgically scraped to expose the subendothelial matrix (Fig. 3a). The intact and damaged artery samples were subsequently incubated with fluorescently labelled PNPs for 30 s followed by repeated PBS washes. The resulting arterial cross-sections and *en face* visualizations revealed that the denuded artery was more prone to PNP adhesion than the intact artery (Fig. 3b, c). In Fig. 3c, it can also be observed that PNPs bind preferentially to the edges of the intact artery, where subendothelium was exposed upon tissue incision. This selective PNP adhesion was further validated in a rat model of angioplasty-induced arterial injury. Pharmacokinetic analyses and biodistribution studies showed that >90% of the PNPs were distributed to tissues 30 min after intravenous injection, with liver and spleen being the primary distribution organs (Extended Data Fig. 6a, b). A comprehensive blood chemistry panel analysis revealed that the PNPs did not inflict observable adverse effects in the rats (Extended Data Fig. 6c). Selective particle binding to the denuded artery was observed upon examination of the aortic branch 2 h after administration of PNP (Fig. 3d and Extended Data Fig. 7). The PNPs were localized on the luminal side above the smooth muscle layer (Fig. 3e), and retention at the injury site lasted for at least 5 days (Fig. 3f). In a rat model of coronary restenosis, therapeutic relevance of platelet-mimicking delivery was examined using docetaxel-loaded PNPs (PNP-Dtxl) (Extended Data Fig. 8). PNP-Dtxl treatment on day 0 and 5 at 0.3 mg per kg body weight (mg kg^{-1}) of docetaxel dosing potentially inhibited neointima growth in balloon-denuded rats as evidenced by the arterial cross-sections collected on day 14 (Fig. 3g, h and Extended Data Fig. 9). To evaluate the vascular remodelling quantitatively, intima-to-media ratio (I/M) and luminal obliteration were calculated. Compared to free docetaxel, which resulted in an I/M of 0.76 ± 0.18 (mean \pm s.d.) and a luminal obliteration of 33.6%, PNP-Dtxl yielded significantly lower values of 0.18 ± 0.06 and 8.0%, respectively ($P \leq 0.0001$) (Fig. 3i, j). These results demonstrate the benefit of PNP-directed delivery in improving drug localization to diseased vasculatures.

We further examined the therapeutic potential of PNPs against platelet-adhering pathogens. Opportunistic bacteria, including several

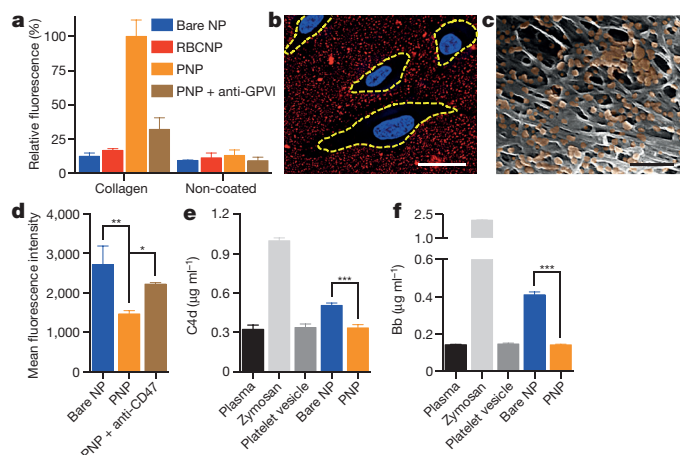
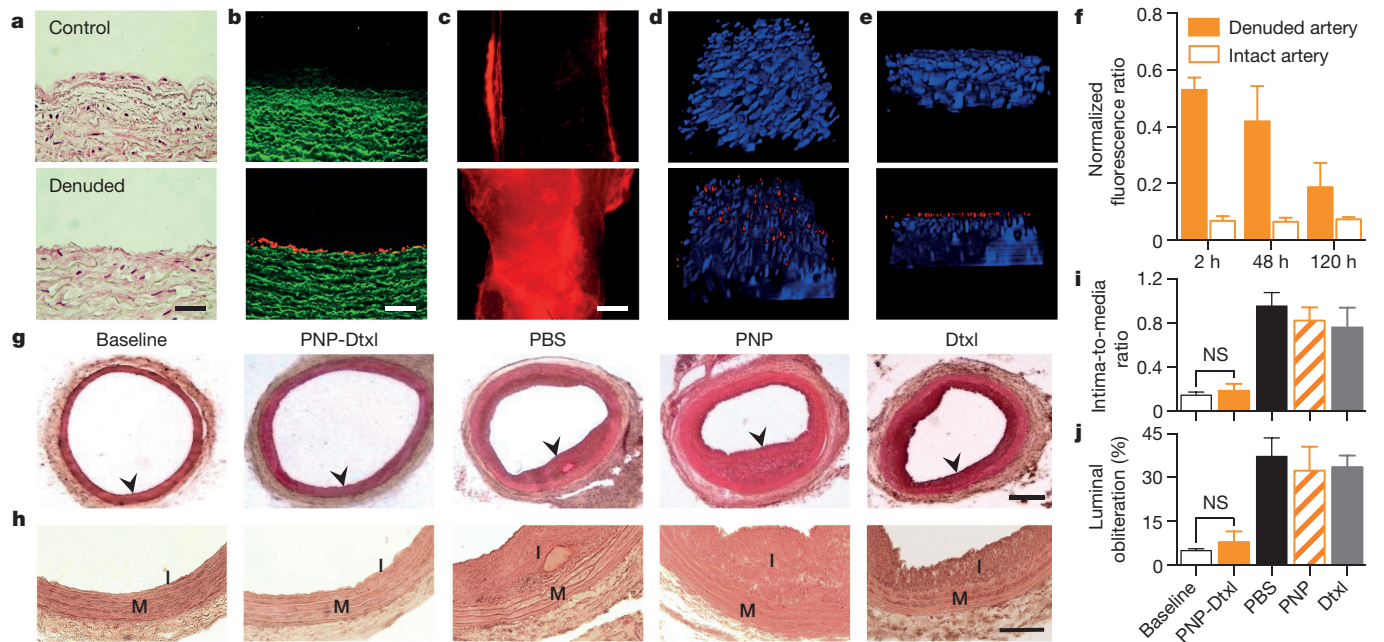
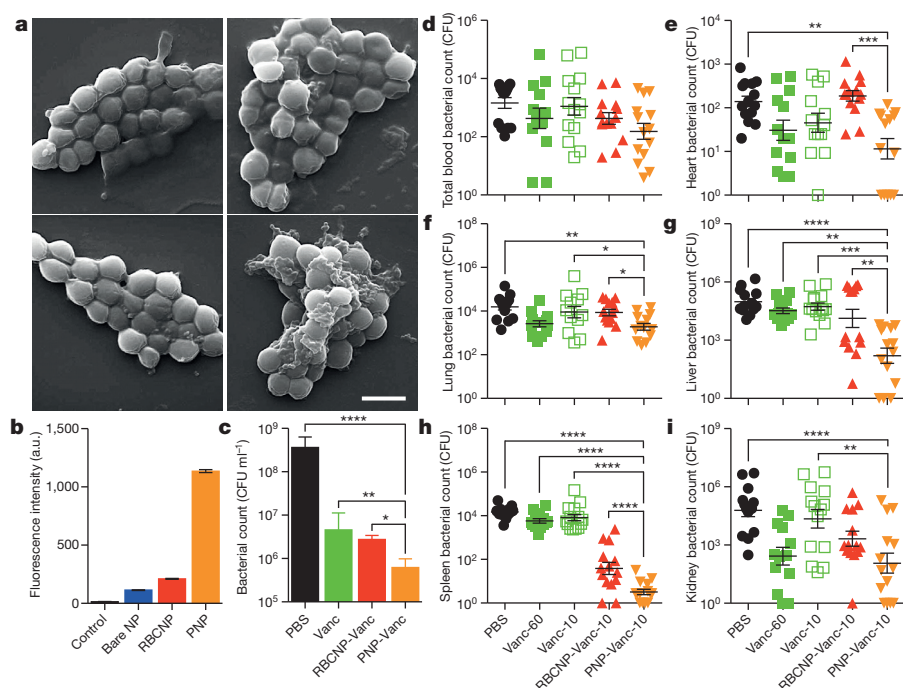


Figure 2 | Collagen binding and immunocompatibility. **a**, Fluorescence quantification of nanoparticle retention on collagen-coated and non-coated plates ($n = 6$). **b**, Localization of PNPs (stained in red) on collagen-coated tissue culture slides seeded with HUVECs (nuclei stained in blue). Cellular periphery is outlined based on cytosolic staining. Scale bar, 10 μm . **c**, A pseudocoloured SEM image of the extracellular matrix of a decellularized human umbilical cord artery after PNP incubation (PNPs coloured in orange). Scale bar, 500 nm. **d**, Flow cytometric analysis of nanoparticle uptake by human THP-1 macrophage-like cells ($n = 3$). **e**, **f**, Classical complement activation measured by C4d split products (**e**) and alternative complement activation measured by Bb split products (**f**) for bare NPs, platelet vesicles, and PNPs in autologous human plasma ($n = 4$). Zymosan and untreated plasma are used as positive and negative controls, respectively. All bars represent means \pm s.d. * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$.



strains of staphylococci and streptococci, exploit platelets by both direct and indirect adherence mechanisms for tissue localization and immune evasion⁶. To demonstrate that PNPs can exploit the inherent bacterial adherence mechanism for targeted antibiotics delivery, MRSA252, a strain of methicillin-resistant *Staphylococcus aureus* expressing a serine-rich adhesin for platelets (SraP)²⁸, was used as a

model pathogen for particle adherence study. After 10 min of incubation between formalin-fixed MRSA252 and different nanoformulations, the collected bacteria showed preferential binding by PNPs (Fig. 4a), exhibiting a 12-fold increase in PNP retention compared to bare NPs (Fig. 4b and Extended Data Fig. 10). This adherence was membrane-specific as RBCNPs showed lower retention than PNPs. The therapeutic potential



model pathogen for particle adherence study. After 10 min of incubation between formalin-fixed MRSA252 and different nanoformulations, the collected bacteria showed preferential binding by PNPs (Fig. 4a), exhibiting a 12-fold increase in PNP retention compared to bare NPs (Fig. 4b and Extended Data Fig. 10). This adherence was membrane-specific as RBCNPs showed lower retention than PNPs. The therapeutic potential

of PNPs was further evaluated using vancomycin-loaded formulations. In an *in vitro* antimicrobial study, live MRSA252 bacteria were briefly incubated with free vancomycin, vancomycin-loaded RBCNPs (RBCNP-Vanc), or vancomycin-loaded PNPs (PNP-Vanc) followed by a wash and culturing in fresh media. The PNP-Vanc formulation showed statistically significant improvement in MRSA252 reduction that corroborates the targeting effect of the particles (Fig. 4c). An *in vivo* antimicrobial efficacy study was further conducted using a mouse model of systemic MRSA252 infection. Mice systemically challenged with 6×10^6 colony-forming units (CFU) of MRSA252 received once daily intravenous treatment of free vancomycin, RBCNP-Vanc, or PNP-Vanc for 3 days at 10 mg kg^{-1} of vancomycin. A control group of high-dose vancomycin treatment in which infected mice received free vancomycin at 30 mg kg^{-1} twice daily was conducted in parallel. 24 h after the last treatment, bacterial enumeration at the primary infection organs showed that the PNP-Vanc resulted in the lowest mean bacterial counts across all organs (Fig. 4d–i). Statistical analyses revealed significance between PNP-Vanc and free vancomycin at equivalent dosage in the lung, liver, spleen and kidney. In comparison to free vancomycin at sixfold the dosage, PNP-Vanc showed significantly better antimicrobial efficacy in the liver and spleen and was at least as effective in the blood, heart, lung and kidney. Notably, compared to RBCNP-Vanc, PNP-Vanc showed significantly higher potency in the heart, lung, liver and spleen, reflecting membrane-specific modulation of nanoparticle performance. The study validates the feasibility of harnessing biomembrane interfaces to improve infectious disease treatment.

The vast medical relevance of platelets has inspired many platelet-mimicking systems that target dysfunctional vasculature in cardiovascular diseases^{8,9}, traumas^{10,11,13}, cancers¹² and acute inflammations²⁹. The present PNP platform exploits platelet membrane in its entirety to enable biomimetic interactions with proteins, cells, tissues and micro-organisms. Towards translation, the platform would benefit from existing infrastructures and logistics for transfusion medicine, polymeric nanotherapeutics and cell-derived pharmaceuticals. Previous work on the cell membrane cloaking approach demonstrated high cloaking efficiency³⁰ and viable storage¹⁸ upon platform optimization (Extended Data Fig. 2f–h). By employing large-scale purification and dispersion techniques commonly applied to biologics, reliable platelet membrane derivation and PNP production can be envisioned.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 November 2014; accepted 12 August 2015.

Published online 16 September 2015.

- Pelaz, B. *et al.* Interfacing engineered nanoparticles with biological systems: anticipating adverse nano-bio interactions. *Small* **9**, 1573–1584 (2013).
- Salvati, A. *et al.* Transferrin-functionalized nanoparticles lose their targeting capabilities when a biomolecule corona adsorbs on the surface. *Nature Nanotechnol.* **8**, 137–143 (2013).
- Tenzen, S. *et al.* Rapid formation of plasma protein corona critically affects nanoparticle pathophysiology. *Nature Nanotechnol.* **8**, 772–781 (2013).
- Born, G. V. & Cross, M. J. The aggregation of blood platelets. *J. Physiol. (Lond.)* **168**, 178–195 (1963).
- Kieffer, N. & Phillips, D. R. Platelet membrane glycoproteins: functions in cellular interactions. *Annu. Rev. Cell Biol.* **6**, 329–357 (1990).
- Fitzgerald, J. R., Foster, T. J. & Cox, D. The interaction of bacterial pathogens with platelets. *Nature Rev. Microbiol.* **4**, 445–457 (2006).
- Yeaman, M. R. Platelets in defense against bacterial pathogens. *Cell. Mol. Life Sci.* **67**, 525–544 (2010).

- Peters, D. *et al.* Targeting atherosclerosis by using modular, multifunctional micelles. *Proc. Natl Acad. Sci. USA* **106**, 9815–9819 (2009).
- Chan, J. M. *et al.* Spatiotemporal controlled delivery of nanoparticles to injured vasculature. *Proc. Natl Acad. Sci. USA* **107**, 2213–2218 (2010).
- Bertram, J. P. *et al.* Intravenous hemostat: nanotechnology to halt bleeding. *Sci. Transl. Med.* **1**, 11ra22 (2009).
- Modery-Pawlowski, C. L. *et al.* Approaches to synthetic platelet analogs. *Biomaterials* **34**, 526–541 (2013).
- Simberg, D. *et al.* Biomimetic amplification of nanoparticle homing to tumors. *Proc. Natl Acad. Sci. USA* **104**, 932–936 (2007).
- Anselmo, A. C. *et al.* Platelet-like nanoparticles: mimicking shape, flexibility, and surface biology of platelets to target vascular injuries. *ACS Nano* **8**, 11243–11253 (2014).
- Olsson, M., Bruhns, P., Frazier, W. A., Ravetch, J. V. & Oldenburg, P. A. Platelet homeostasis is regulated by platelet expression of CD47 under normal conditions and in passive immune thrombocytopenia. *Blood* **105**, 3577–3582 (2005).
- Sims, P. J., Rollins, S. A. & Wiedmer, T. Regulatory control of complement on blood platelets. Modulation of platelet procoagulant responses by a membrane inhibitor of the C5b-9 complex. *J. Biol. Chem.* **264**, 19228–19235 (1989).
- Nieswandt, B. & Watson, S. P. Platelet-collagen interaction: is GPVI the central receptor? *Blood* **102**, 449–461 (2003).
- Hu, C. M. *et al.* Erythrocyte membrane-camouflaged polymeric nanoparticles as a biomimetic delivery platform. *Proc. Natl Acad. Sci. USA* **108**, 10980–10985 (2011).
- Hu, C. M., Fang, R. H., Copp, J., Luk, B. T. & Zhang, L. A biomimetic nanosponge that absorbs pore-forming toxins. *Nature Nanotechnol.* **8**, 336–340 (2013).
- Hu, C. M., Fang, R. H., Luk, B. T. & Zhang, L. Nanoparticle-detained toxins for safe and effective vaccination. *Nature Nanotechnol.* **8**, 933–938 (2013).
- Gachet, C. *et al.* Alpha IIb beta 3 integrin dissociation induced by EDTA results in morphological changes of the platelet surface-connected canalicular system with differential location of the two separate subunits. *J. Cell Biol.* **120**, 1021–1030 (1993).
- Luk, B. T. *et al.* Interfacial interactions between natural RBC membranes and synthetic polymeric nanoparticles. *Nanoscale* **6**, 2730–2737 (2014).
- Hughes, C. E. *et al.* CLEC-2 activates Syk through dimerization. *Blood* **115**, 2947–2955 (2010).
- Kalluri, R. Basement membranes: structure, assembly and role in tumour angiogenesis. *Nature Rev. Cancer* **3**, 422–433 (2003).
- Rodriguez, P. L. *et al.* Minimal “Self” peptides that inhibit phagocytic clearance and enhance delivery of nanoparticles. *Science* **339**, 971–975 (2013).
- Law, S. K. A. & Dodds, A. W. The internal thioester and the covalent binding properties of the complement proteins C3 and C4. *Protein Sci.* **6**, 263–274 (1997).
- Terstappen, L. W. M. M., Nguyen, M., Lazarus, H. M. & Medof, M. E. Expression of the DAF (CD55) and CD59 antigens during normal hematopoietic cell differentiation. *J. Leukoc. Biol.* **52**, 652–660 (1992).
- Andersen, A. J., Hashemi, S. H., Andresen, T. L., Hunter, A. C. & Moghimi, S. M. Complement: alive and kicking nanomedicines. *J. Biomed. Nanotechnol.* **5**, 364–372 (2009).
- Siboo, I. R., Chambers, H. F. & Sullam, P. M. Role of SraP, a Serine-Rich Surface Protein of *Staphylococcus aureus*, in binding to human platelets. *Infect. Immun.* **73**, 2273–2280 (2005).
- Kamaly, N. *et al.* Development and *in vivo* efficacy of targeted polymeric inflammation-resolving nanoparticles. *Proc. Natl Acad. Sci. USA* **110**, 6506–6511 (2013).
- Hu, C. M. *et al.* ‘Marker-of-self’ functionalization of nanoscale particles through a top-down cellular membrane coating approach. *Nanoscale* **5**, 2664–2668 (2013).

Acknowledgements This work is supported by the National Institutes of Health under Award Numbers R01DK095168 (L.Z.), R01HL108735 (S.C.) and R01EY25090 (K.Z.), and partially by the Defense Threat Reduction Agency Joint Science and Technology Office for Chemical and Biological Defense under Grant Number HDTRA1-14-1-0064 (L.Z.). R.H.F. is supported by a National Institutes of Health R25CA153915 training grant from the National Cancer Institute.

Author Contributions C.-M.J.H., R.H.F., K.-C.W., B.T.L., K.Z., S.C. and L.Z. conceived and designed the experiments; C.-M.J.H., R.H.F., K.-C.W., B.T.L., S.T., D.D., P.N., P.A., C.H.W., A.V.K., C.C., M.R., V.Q., S.H.P., J.Z., W.S., F.M.H., T.C.C. and W.G. performed all the experiments. The manuscript was written by C.-M.J.H., R.H.F., B.T.L., W.G. and L.Z. All authors discussed the results and reviewed the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.Z. (zhang@ucsd.edu), S.C. (shuchien@ucsd.edu) or K.Z. (kang.zhang@gmail.com).

METHODS

Human platelet isolation and platelet membrane derivation. Human type O⁻ blood anti-coagulated with 1.5 mg ml⁻¹ EDTA was purchased from BioreclamationIVT and processed for platelet collection approximately 16 h after blood collection. Unless otherwise stated, platelets derived from this commercial blood source were used in this study. Fresh human type O⁻ blood was also collected with dipotassium EDTA-treated or lithium heparin-treated blood collection tubes (Becton, Dickinson and Company) under the approval of the Institutional Review Board (IRB) at the University of California, San Diego, USA. Patients consented to use of their blood samples for this study before collection. The freshly drawn blood was processed for platelet collection approximately 30 min after blood draw. In addition, unexpired (in-dated) human type O⁻ platelet rich plasma (PRP) in acid-citrate-dextrose (ACD) was purchased from the San Diego Blood Bank. Samples not originally drawn in EDTA were adjusted to a concentration of 5 mM EDTA before platelet collection. To isolate platelets, the blood and plasma samples were centrifuged at 100g for 20 min at room temperature to separate red blood cells and white blood cells. The resulting PRP was then centrifuged at 100g for 20 min to remove remaining blood cells. PBS buffer containing 1 mM of EDTA and 2 μ M of prostaglandin E1 (PGE1, Sigma Aldrich) was added to the purified PRP to prevent platelet activation. Platelets were then pelleted by centrifugation at 800g for 20 min at room temperature, after which the supernatant was discarded and the platelets were resuspended in PBS containing 1 mM of EDTA and mixed with protease inhibitor tablets (Pierce). 1.5 ml aliquots of platelet solution containing $\sim 3 \times 10^9$ platelets were prepared and used to cloak 1 mg of PLGA nanoparticles.

Platelet membrane was derived by a repeated freeze-thaw process. Aliquots of platelet suspensions were first frozen at -80°C , thawed at room temperature, and pelleted by centrifugation at 4,000g for 3 min. After three repeated washes with PBS solution mixed with protease inhibitor tablets, the pelleted platelet membranes were suspended in water and sonicated in a capped glass vial for 5 min using a Fisher Scientific FS30D bath sonicator at a frequency of 42 kHz and a power of 100 W. The presence of platelet membrane vesicles was verified by size measurement using dynamic light scattering (DLS) and morphological examination by transmission electron microscopy (TEM).

Platelet membrane-cloaked nanoparticle (PNP) preparation and characterization. 100 nm polymeric cores were prepared using 0.67 dl g⁻¹ carboxyl-terminated 50:50 poly(lactic-co-glycolic) acid (PLGA) (LACTEL Absorbable Polymers) in a nanoprecipitation process. 1 ml of 10 mg ml⁻¹ PLGA solution in acetone was added dropwise to 3 ml of water. For fluorescently labelled nano formulations, 1,1'-diiododecyl-3,3',3'-tetramethylindodicarbocyanine perchlorate (DiD, excitation = 644 nm/emission = 665 nm, Life Technologies) was loaded into the polymeric cores at 0.1 wt%. The mixture was then stirred in open air for 1 h and placed in vacuum for another 3 h. The resulting nanoparticle solution was filtered with 10 kDa MWCO Amicon Ultra-4 Centrifugal Filters (Millipore). Platelet membrane cloaking was then accomplished by dispersing and fusing platelet membrane vesicles with PLGA particles by sonication using an FS30D bath sonicator at a frequency of 42 kHz and a power of 100 W for 2 min. The size and the surface zeta potential of replicate PNP samples ($n = 3$) were obtained by dynamic light scattering (DLS) measurements using a Malvern ZEN 3600 Zetasizer. PBS stability was examined by mixing 1 mg ml⁻¹ of PNPs in water with 2 \times PBS at a 1:1 volume ratio. Storbility of PNPs was examined by suspending PNPs in 10% sucrose. The nanoparticle solutions were subject to either a freeze-thaw cycle or lyophilization followed by resuspension. The resulting particle solution was then monitored for particle size using DLS. The structure of PNPs was examined with TEM after negative staining with 1 wt% uranyl acetate using an FEI 200 kV Sphera microscope. RBCNPs were prepared using the same polymeric cores and RBC membranes of equivalent total surface area to the platelet membranes following a previously described protocol¹⁶. The RBCNPs were characterized using DLS and had similar size and zeta potential as the PNPs.

Docetaxel-loaded PLGA nanoparticle cores were prepared by a nanoprecipitation process. 10 wt% docetaxel was added to 5 mg PLGA in acetone and precipitated dropwise into 3 ml water. The solvent was evaporated as described above and free docetaxel was removed by repeated wash steps. Vancomycin-loaded nanoparticles were synthesized using a double emulsion process. The inner aqueous phase consisted of 25 μ l of vancomycin (Sigma Aldrich) dissolved in 1 M NaOH at 200 mg ml⁻¹. The outer phase consisted of 500 μ l of PLGA polymer dissolved in dichloromethane at 50 mg ml⁻¹. The first emulsion was formed by sonication at 70% power pulsed (2 s on/1 s off) for 2 min on a Fisher Scientific 150E Sonic Dismembrator. The resulting emulsion was then emulsified in aqueous solution under the same dispersion setting. The final water/oil/water emulsion was added to 10 ml of water and the solvent was evaporated in a fume hood under gentle stirring for 3 h. The particles were collected by centrifugation at 80,000g in a Beckman Coulter Optima L-90K Ultracentrifuge. The particles were washed

and resuspended in water. Upon preparation of drug-loaded PLGA cores, cell membrane coating was performed by adding the appropriate surface area equivalent of either platelet or red blood cell membrane followed by 3 min of sonication in a Fisher Scientific FS30D Bath Sonicator. Particle size, polydispersity (PDI), and surface zeta potential were characterized using DLS. Drug loading yield and release rate of replicate samples ($n = 3$) were quantified by high performance liquid chromatography (HPLC). Drug release was determined by dialyzing 500 μ l of particle solution at a concentration of 2.67 mg ml⁻¹ in PBS using 3.5 K MWCO Slide-A-Lyzers (Thermo Scientific).

Examination of platelet membrane proteins. PNPs were purified from unbound proteins or membrane fragments by centrifugation at 16,000g in 10% sucrose. Platelet-rich plasma, platelets, platelet membrane vesicles, and PNPs were then normalized to equivalent overall protein concentration using a Pierce BCA Protein Assay Kit (Life Technologies). To examine the effect of different platelet derivation protocols on the membrane protein expression, platelets collected from commercial blood anti-coagulated in EDTA, freshly drawn blood anti-coagulated in EDTA or heparin, and transfusion-grade PRP in ACD were prepared in parallel. All platelets were processed using the aforementioned platelet membrane derivation protocol for PNP preparation. The samples containing equivalent total proteins were then lyophilized, prepared in lithium dodecyl sulfate (LDS) sample loading buffer (Invitrogen), and separated on a 4–12% Bis-Tris 17-well minigel in MOPS running buffer using a Novex Xcell SureLock Electrophoresis System (Life Technologies). Identification of key membrane proteins by western blotting was performed using primary antibodies including mouse anti-human CD47 (eBioscience, B6H12), mouse anti-human CD55 (Biolegend, JS11), mouse anti-human CD59 (Biolegend, p282 (H19)), mouse anti-human integrin α IIb subunit (Biolegend, HIP8), rat anti-human integrin α 2 subunit (R&D Systems, 430907), rabbit anti-human integrin α 5 subunit (Abgent, AP12204c), mouse anti-human integrin α 6 subunit (Abgent, AM1828a), mouse anti-human integrin β 1 subunit (R&D Systems, 4B7R), mouse anti-human integrin β 3 subunit (Biolegend, VI-PL2), mouse anti-human GPIIb γ (R&D Systems, 486805), mouse anti-human GPIV (R&D Systems, 877346), mouse anti-human GPV (Santa Cruz Biotech, G-11), rat anti-human GPVI (EMD Millipore, 8E9), rabbit anti-human GPXI (Santa Cruz Biotech, A-9), and mouse anti-human CLEC-2 (Genetex, 8J24). A goat anti-mouse IgG-HRP conjugate (Biolegend, Poly4053), a goat anti-rat IgG-HRP conjugate (Biolegend, Poly4054), or a donkey anti-rabbit IgG-horseradish peroxidase (HRP) conjugate (Biolegend, Poly4064) was used for secondary staining based on the isotype of the primary antibody. MagicMark XP western protein standard (Invitrogen) was used as a molecular weight ladder. The nitrocellulose membrane was then incubated with ECL western blotting substrate (Pierce) and developed with the Mini-Medical/90 Developer (ImageWorks).

Examination of protein sidedness on PNPs. For immunogold staining, a drop of the PNP solution (1 mg ml⁻¹) was deposited onto a glow-discharged carbon-coated grid. The grid was then washed 3 times with PBS, blocked with 1% BSA for 15 min, and stained with 0.5 mg ml⁻¹ of anti-CD47 targeted to either the intracellular or extracellular domain of the protein. After 1 h of incubation, the samples were rinsed with PBS containing 1% BSA for 6 times and stained with anti-rabbit IgG-gold conjugate (5 nm) solution (Sigma Aldrich) for another hour. After 6 PBS washes, the samples were fixed with 1% glutaraldehyde in PBS for 5 min and washed with water 6 times. The sample grids were subsequently stained with 2% vanadium solution (Abcam) and visualized using an FEI 200 kV Sphera microscope.

For flow cytometric analysis, 2.0 μ m carboxyl-functionalized polystyrene beads at a concentration of 4 wt% (Life Technologies) were functionalized with rabbit N terminus-targeted (extracellular) anti-human CD47 (Aviva Biosystems, ARP63284), rabbit intracellular-domain-targeted anti-CD47 (Genetex, EPR4150(2)), or rabbit anti-ovalbumin (Abcam, ab1221) as a sham antibody by EDC/NHS chemistry. The resulting antibody-modified beads were re-suspended in 100 μ l of DI water. The bead solution was first incubated with 1 mg of bovine serum albumin (BSA, Sigma Aldrich) to block non-specific interactions and then mixed with 1 ml of fluorescently labelled PNPs (200 μ g ml⁻¹). The mixture solution was incubated at room temperature for 2 h and then centrifuged to remove the unbound PNPs. The collected polystyrene beads were then subjected to flow cytometric analysis.

Platelet aggregation assay. Platelets, platelet membrane vesicles, and PNPs of equivalent membrane content were prepared and examined for platelet-activating molecules, including thrombin, ADP, and thromboxane, using a SensoLyte 520 Thrombin Activity Assay Kit (AnaSpec), ADP Colorimetric/Fluorometric Assay Kit (Sigma Aldrich), and Thromboxane B2 (TXB₂) ELISA Kit (Enzo Life Sciences), respectively, based on the manufacturers' instructions. Each sample was assayed in replicate ($n = 3$).

Aggregation of platelets in the presence of PNPs was assessed using a spectrophotometric method. 1 ml aliquot of platelet rich plasma (PRP) was first prepared

from human whole blood with sodium citrate as the anti-coagulant. The plasma was then loaded into a cuvette followed by addition of 500 μl of 2 mg ml^{-1} PNP in PBS solution. As negative and positive controls, the PRP was mixed with 500 μl of PBS or 500 μl of PBS containing 0.5 IU ml^{-1} of human thrombin (Sigma Aldrich), respectively. The cuvettes were immediately placed in a TeCan Infinite M200 reader and monitored for change in absorbance at 650 nm over time, and platelet aggregation was observed based on the reduction of turbidity.

Collagen binding study. Collagen type IV derived from human placenta (Sigma Aldrich) was reconstituted to a concentration of 2.0 mg ml^{-1} in 0.25% acetic acid. 200 μl of the collagen solution was then added to each well of a 96-well assay plate and incubated overnight at 4 °C. Prior to the collagen binding study, the plate was blocked with 2% BSA and washed three times with PBS. For the collagen binding study, 100 μl of 1 mg ml^{-1} DiD-loaded nanoformulations in water were added into replicate wells ($n = 6$) of collagen-coated or non-collagen-coated plates. After 30 s of incubation, the plates were washed three times. Retained nanoparticles were then dissolved with 100 μl of DMSO for fluorescence quantification using a TeCan Infinite M200 reader.

Differential adhesion to endothelial and collagen surfaces. Collagen type IV was coated on 8-well Lab-Tek II chamber slides (Nunc) as described above. The collagen-coated chamber slides were used to seed primary HUVECs obtained from the American Type Culture Collection and cultured in HUVEC Culture Medium (Sigma Aldrich) supplemented with 10% fetal bovine serum for 24 h. The cells were then incubated with 1 mg ml^{-1} DiD-loaded PNP in PBS at 4 °C for 30 s. Next the cells were washed with PBS three times and fixed with tissue fixative (Millipore) for 30 min at room temperature. Fluorescence staining was done with 4',6-diamidino-2-phenylindole (DAPI, Life Technologies) for the nuclei and 22-(*n*-(7-nitrobenz-2-oxa-1,3-diazol-4-yl)amino)-23,24-bisnor-5-cholesterol-3 β -ol (NBD cholesterol, Life Technologies) for the cytosol before mounting the cells in ProLong Gold antifade reagent (Life Technologies) and imaged using a DeltaVision deconvolution scanning fluorescence microscope. z-stacks were collected at 0.25 μm intervals over 10 μm . The images were deconvolved and superimposed. DiD fluorescence signal over collagen and endothelial surfaces as defined by the boundaries of NBD fluorescence were analysed using ImageJ. PNP retention over collagen and endothelial surfaces was quantified based on distinct images ($n = 10$) in which the average fluorescence per unit area was analysed.

Cellular uptake study with macrophage-like cells. THP-1 cells were obtained directly from the American Type Culture Collection and used without further authentication or testing for mycoplasma contamination. The cells were maintained in RPMI 1640 media (Life Technologies) supplemented with 10% FBS (Sigma Aldrich). THP-1 cells were differentiated in 100 ng ml^{-1} phorbol myristate acetate (PMA, Sigma Aldrich) for 48 h and differentiation was visually confirmed by cellular attachment to Petri dishes. For the cellular uptake study, the differentiated macrophage-like cells were incubated in replicate wells ($n = 3$) with DiD-loaded PNP, anti-CD47 blocked PNP, and bare NP at 100 $\mu\text{g ml}^{-1}$ in culture media. After 30 min of incubation at 37 °C, the macrophage-like cells were scraped off the Petri dish and washed three times in PBS to remove non-internalized particles. Flow cytometry was performed to examine nanoparticle uptake by the macrophage-like cells. All flow cytometry studies were conducted on a FACSCanto II flow cytometer (BD Biosciences) and the data was analysed using FlowJo software from Tree Star. Statistical analysis was performed based on a two-tailed, unpaired *t*-test.

Complement activation study. To assess complement system activation, two complement split products (C4d and Bb) were analysed using enzyme-linked immunosorbent assay kits (Quidel Corporation). The nanoparticles were incubated in replicate aliquots ($n = 4$) of human serum at a volume ratio of 1:5 in a shaking incubator (80 r.p.m.) at 37 °C for 1 h. The reaction was then stopped by adding 60 volumes of PBS containing 0.05% Tween-20 and 0.035% ProClin 300. Complement system activation of the nanoparticles was assayed following the manufacturer's instructions, and zymosan was used as a positive control.

PNP adherence to human carotid artery. Human umbilical cord was collected under the approval of the Institutional Review Board (IRB) at the University of California, San Diego, USA, and human carotid arteries were collected under the approval of the IRB at the University of Southern California, USA. Patients consented to use of their samples for this study before collection. To derive decellularized arterial extracellular matrix (ECM), human arteries were carefully dissected from the umbilical cord and removed from the surrounding Wharton's jelly, and subsequently incubated in 2% sodium dodecyl sulfate (SDS, Sigma Aldrich) for 72 h. The decellularized tissue was then rinsed with PBS and incubated in PBS solution containing 200 $\mu\text{g ml}^{-1}$ PNP for 30 s. The sample was then transferred to PBS solution and rinsed extensively before examination by scanning electron microscopy (SEM). A control decellularized arterial tissue sample without PNP incubation was prepared and visualized for comparison.

To examine PNP binding on denuded vascular walls, approximately 2 mm thick fresh human carotid artery sections were dissected and placed in normal saline on ice and transported immediately to the laboratory for a PNP binding study. To create the vascular characteristics of damaged arteries, an excised artery sample was surgically scraped on its luminal side with forceps to remove the endothelial layer. Successful denudation was confirmed by microscopy visualization. Prior to the nanoparticle binding experiment, both damaged and non-damaged artery samples were rinsed with PBS solution. The PNP binding experiment was performed by incubating the arterial samples in PBS solution containing 200 $\mu\text{g ml}^{-1}$ of DiD-loaded PNP for 30 s. The samples were then transferred to PBS solution and rinsed extensively before visualization by fluorescence microscopy. Endogenous tissue components such as collagen and elastin were identified based on their autofluorescence, which excites and emits maximally at $\sim 300 - 500$ nm and was captured using a FITC filter. DiD fluorescence was captured using a Cy5 filter to examine the deposition of PNP. The arterial samples were imaged by a cross-sectional view of a histological section and a top-down view on the luminal side. The images were normalized to a reference illumination image for proper comparison.

Pharmacokinetics, biodistribution and safety of PNP in a rat model of angioplasty-induced arterial denudation. All animal experiments were performed in accordance with NIH guidelines and approved by the Animal Care Committee of the University of California, San Diego. For the pharmacokinetics study, adult male Sprague-Dawley rats weighing 300–350 g (Harlan Laboratories) were administered with DiD-labelled PNP and their blood was collected at specific time points via tail-vein blood sampling for fluorescence quantification. For the safety study, rats were injected with 1 ml of 5 mg ml^{-1} of PNP on day 0 and day 5 followed by blood collection on day 10 for comprehensive metabolic panel analysis. Rats receiving equivalent PBS injections were prepared as a control.

For the biodistribution and vasculature-targeting studies, adult male Sprague-Dawley rats weighing 300–350 g (Harlan Laboratories) were subjected to carotid balloon injury. In brief, the animals were anaesthetized with intraperitoneal ketamine (Pfizer) at 100 mg kg^{-1} and xylazine (Lloyd Laboratories) at 10 mg kg^{-1} . A ventral mid-line incision (~ 2 cm) was made in the neck, and the left common carotid artery and carotid bifurcation were exposed by blunt dissection. Proximal of left carotid artery, inner carotid artery and external carotid artery were temporarily clamped to avoid excessive blood loss during the induction of the 2F Fogarty arterial embolectomy catheter (Edwards Lifesciences). The catheter was introduced into the left carotid artery through an arteriotomy on the external carotid artery. The catheter was slowly inflated to a determined volume (0.02 ml) and withdrawn with rotation for 3 times to denude the endothelium. The wound was later closed with 4-0 sutures.

After the wound closure, rats were injected intravenously with 1 ml of 5 mg ml^{-1} DiD-loaded PNP in 10% sucrose. At specified time points after the injection, animals were euthanized by CO₂ inhalation. After perfusion with PBS, organs including heart, lung, liver, spleen, kidney, gut, blood, and aortic branch including both the left and right carotid arteries were carefully collected and homogenized for biodistribution analysis. The overall PNP distribution at the aortic branch was visualized using a Keyence BZ-X700 fluorescence microscope. To examine the local distribution of PNP, damaged and undamaged arteries were cut longitudinally and stained with DAPI solution. *En face* examination was done on the luminal surfaces of denuded and intact areas for the binding of PNP. Sequences of images along the z-axis (0.5 μm per section) from the intima to media layers of the carotid arteries were acquired with an Olympus ix81 fluorescence microscope. The 3D reconstruction of the arterial wall from multisectional images was performed using ImageJ. To analyse the PNP retention, damaged and undamaged arteries collected at specified time points were homogenized, and their respective fluorescence was normalized to the liver fluorescence for comparison. All replicates represent different rats subjected to the same treatment ($n = 6$).

Treatment of experimental coronary restenosis. Sprague-Dawley rats after angioplasty-induced arterial denudation were randomly placed into groups. The nanoparticle treatment group was injected intravenously with 1 ml of 5 mg ml^{-1} docetaxel-loaded PNP in 10% sucrose at a docetaxel dose of 0.3 mg kg^{-1} on day 0 and day 5. As controls, animals receiving PBS, free docetaxel, and empty PNP were prepared. On day 14, animals were euthanized with an overdose of a ketamine-xylazine cocktail and perfused with PBS and 4% paraformaldehyde (PFA) at a pressure of 120–140 mm Hg. Segments of left and right carotid arteries were carefully dissected out, and the PFA-fixed carotid arteries were embedded with Tissue-Tek OCT compound (VWR International) in a tissue base mould and slowly submerged into pre-chilled 2-methyl butane until frozen completely. The frozen tissue block was then immediately stored at -80 °C until sectioning. Serial sectioning (15 μm per section) was performed with a Cryotome cryostat machine (Leica), and the tissue sections were placed on polylysine-treated glass slides. Tissue sections on slides were dried at room temperature for 30 min before stain-

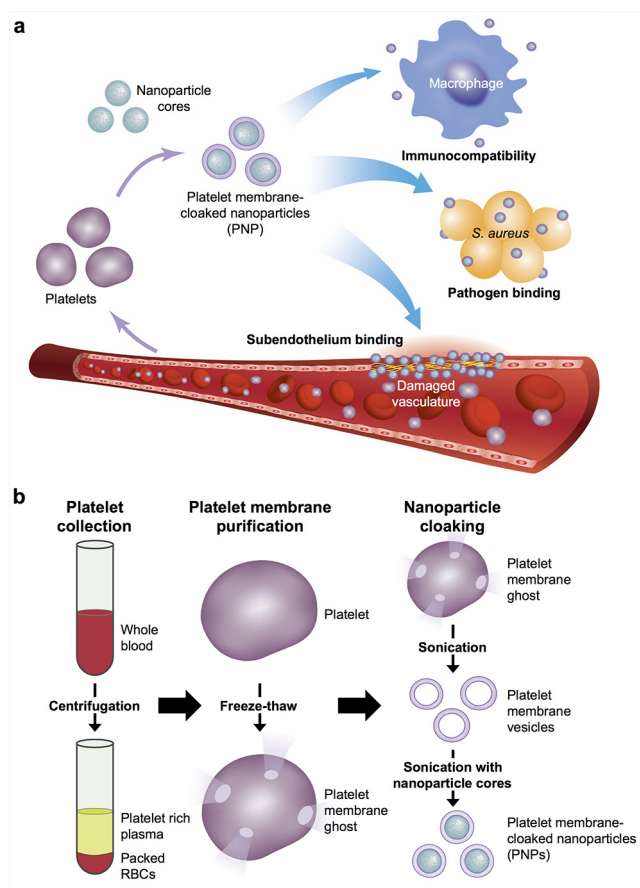
ing. For immunohistochemistry, frozen sections on slides were first washed with PBS to remove residual OCT medium and then subjected to standard haematoxylin and eosin (H&E) staining. Areas of intima and media were analysed using Image J. Luminal obliteration is defined as the intima area/the area within the internal elastic lamina. Statistical analysis was performed using one-way ANOVA. No statistical methods were used to predetermine sample size. Studies were done in a non-blinded fashion. All replicates represent different rats subjected to the same treatment ($n = 6$).

***Staphylococcus aureus* (MRSA252) bacteria adherence study.** MRSA252 obtained from the American Type Culture Collection was cultured on tryptic soy broth (TSB) agar (Becton, Dickinson and Company) overnight at 37 °C. A single colony was inoculated in TSB medium at 37 °C in a rotary shaker. Overnight culture was refreshed in TSB medium at a 1:100 dilution at 37 °C under shaking for another 3 h until the OD₆₀₀ of the culture medium reached approximately 1.0 (logarithmic growth phase). The bacteria were harvested by centrifugation at 5,000g for 10 min and then washed with sterile PBS twice and then fixed with 10% formalin for 1 h. The fixed bacteria were washed with sterile PBS and suspended in 10% sucrose to a concentration of 1×10^8 CFU ml⁻¹. For the nanoparticle adhesion study, aliquots of 0.8 ml of 1×10^8 CFU ml⁻¹ MRSA252 were mixed with 1.2 ml of 200 µg ml⁻¹ DiD-loaded PNPs, RBCNPs, or bare NPs in 10% sucrose for 10 min at room temperature. The bacteria were then isolated from unbound nanoparticles by repeated centrifugal washes in sucrose solution at 5,000g. The purified bacteria were then suspended in 10% sucrose for replicate measurements ($n = 3$) by flow cytometric analysis and SEM imaging.

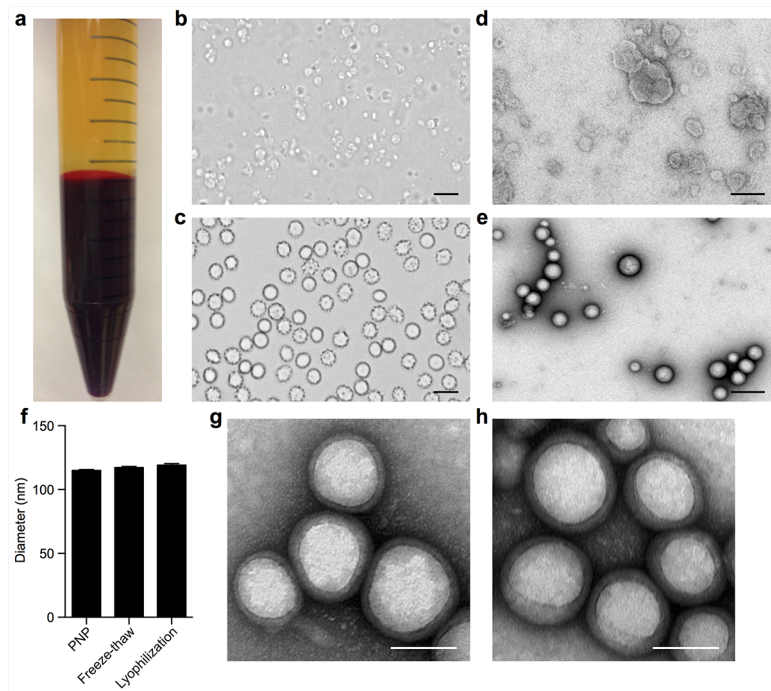
Antimicrobial efficacy study. For the *in vitro* antimicrobial efficacy study, 5×10^6 CFU of MRSA252 was mixed with 500 µl of 20 mg ml⁻¹ nanoparticles (4 wt% vancomycin loading) in saline. As controls, equivalent amounts of bacteria were incubated in either PBS or free vancomycin (0.8 mg ml⁻¹). After 10 min of incubation, bacteria were isolated from the solution by centrifugation at 2,500g for 5 min. The collected bacteria pellet was resuspended with 500 µl of TSB culture medium and incubated for 5 h. The resulting samples were serially diluted in PBS and spotted on TSB agar plates. After 24 h of culturing, the colonies were counted

to determine the bacteria count in each sample. Replicates represent separate bacterial aliquots incubated with the same formulation ($n = 3$).

For the *in vivo* antimicrobial efficacy study, vancomycin-loaded PNPs (PNP-Vanc) and vancomycin-loaded RBCNPs (RBCNP-Vanc) were suspended in 10% sucrose solution at 31.25 mg ml⁻¹ (4 wt% vancomycin loading). An equivalent concentration of free vancomycin (1.25 mg ml⁻¹) was also suspended in 10% sucrose. Male CD-1 mice (Harlan Laboratories) weighing ~25 g were challenged intravenously with 6×10^6 CFU of MRSA252 suspended in 100 µl of PBS. 30 min after the bacteria injection, mice were randomly placed into separate groups and injected with 200 µl of PNP-Vanc, RBCNP-Vanc, free vancomycin (daily dosage: 10 mg kg⁻¹ vancomycin), or PBS. To compare to the clinical dosing of vancomycin, a control group treated with twice daily dosing of 30 mg kg⁻¹ free vancomycin was prepared (total daily dosage: 60 mg kg⁻¹ vancomycin). The mice received their corresponding treatments from day 0 to 2. On day 3, blood was collected from the submandibular vein. The mice were then euthanized, perfused with PBS, and their organs were collected. The organs were homogenized using a Biospec Mini Beadbeater in 1 ml of PBS for 1 min, serially diluted in PBS by tenfold, and plated onto agar plates with a spotting volume of 50 µl. After 48 h of culture, bacterial colonies were counted to determine the bacterial load in each organ. Under the given experimental conditions, the detection limit was determined to be approximately 20 CFU per organ. Data points on the *x*-axis represent samples with no detectable bacterial colonies. It was confirmed that samples prepared from unchallenged mice had no detectable colonies. The data was tested for normal distribution using the Shapiro-Wilk test. For blood and heart, which contained non-normal distributions, statistical analysis was performed using Kruskal–Wallis test. For the other organs, in which all groups were normally distributed and variance criteria were met, statistical analysis was performed using one-way ANOVA. Grubbs' test was used to detect and remove statistical outliers. No statistical methods were used to predetermine sample size. Studies were done in a non-blinded fashion. Replicates represent different mice subjected to the same treatment ($n = 14$).

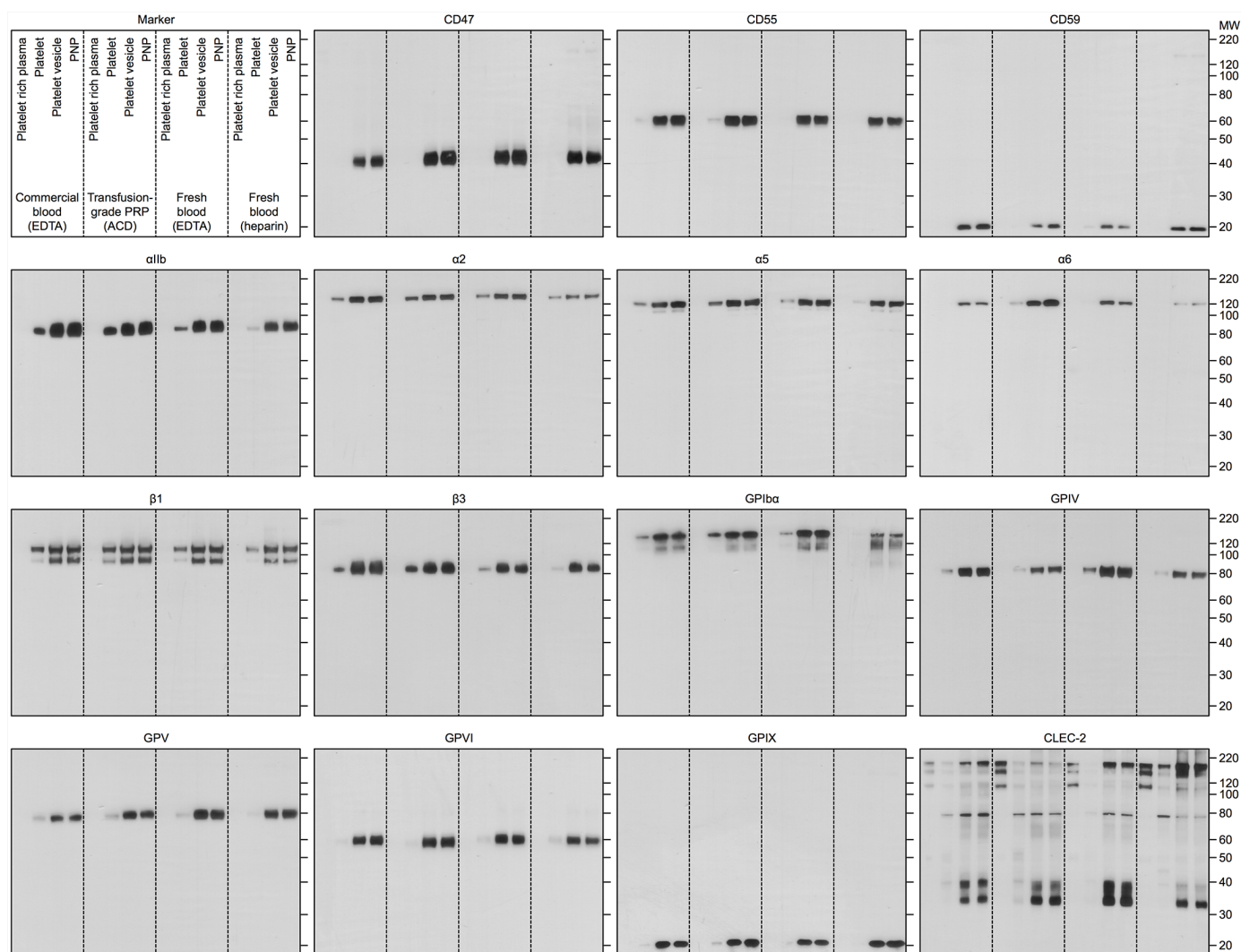


Extended Data Figure 1 | Schematic preparation of PNPs. **a**, Poly(lactic-co-glycolic acid) (PLGA) nanoparticles are enclosed entirely in plasma membrane derived from human platelets. The resulting particles possess platelet-mimicking properties for immunocompatibility, subendothelium binding, and pathogen adhesion. **b**, Schematic depicting the process of preparing PNPs.



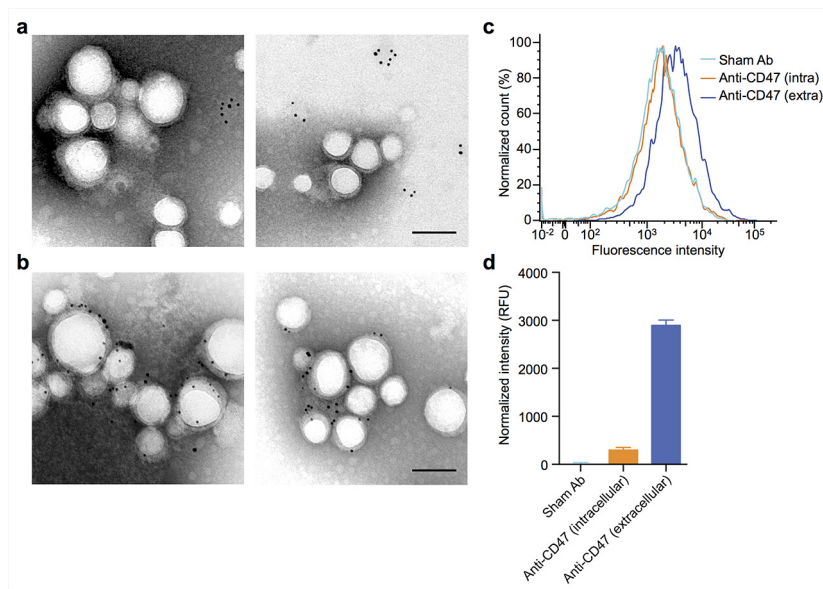
Extended Data Figure 2 | PNP preparation and storage. **a**, Isolation of platelet rich plasma (PRP) was achieved by centrifugation at 100g. PRP was collected from the top layer (yellow) separated from the red blood cells (red, bottom layer). **b**, Collected human platelets under light microscopy, which possess a distinctive morphology from **c**, red blood cells. Scale bars, 10 μ m. **d**, Transmission electron micrographs of platelet membrane vesicles and **e**, PNPs, both of which were negatively stained with 1% uranyl acetate. Scale bars, 200 nm. **f**, Dynamic light scattering measurements of PNPs in 10%

sucrose show that the particles retain their size and stability after a freeze-thaw cycle and re-suspension upon lyophilization ($n = 3$). Bars represent means \pm s.d. **g**, Transmission electron micrograph shows retentions of the core-shell structure of PNPs after a freeze-thaw cycle in 10% sucrose. Scale bar, 100 nm. **h**, Transmission electron micrograph shows retentions of the core-shell structure of PNPs upon resuspension after lyophilization in 10% sucrose. Scale bar, 100 nm.



Extended Data Figure 3 | Overall protein content on PNPs resolved by western blotting. Primary platelet membrane protein/protein subunits including CD47, CD55, CD59, α IIb, α 2, α 5, α 6, β 1, β 3, GPIIb, GPIV, GPV, GPVI, GPIX, and CLEC-2 were monitored in platelet rich plasma, platelets, platelet vesicles, and PNPs. Platelets derived from four different protocols, including commercial blood anti-coagulated in EDTA, freshly drawn blood anti-coagulated in EDTA, freshly drawn blood anti-coagulated in heparin, and

transfusion-grade platelet rich plasma anti-coagulated in acid-citrate-dextrose (ACD), were examined to compare the membrane protein expression. Each sample was normalized to equivalent overall protein content before western blotting. It was observed that the PNP preparation resulted in membrane protein retention and enrichment very similar across the different platelet sources.

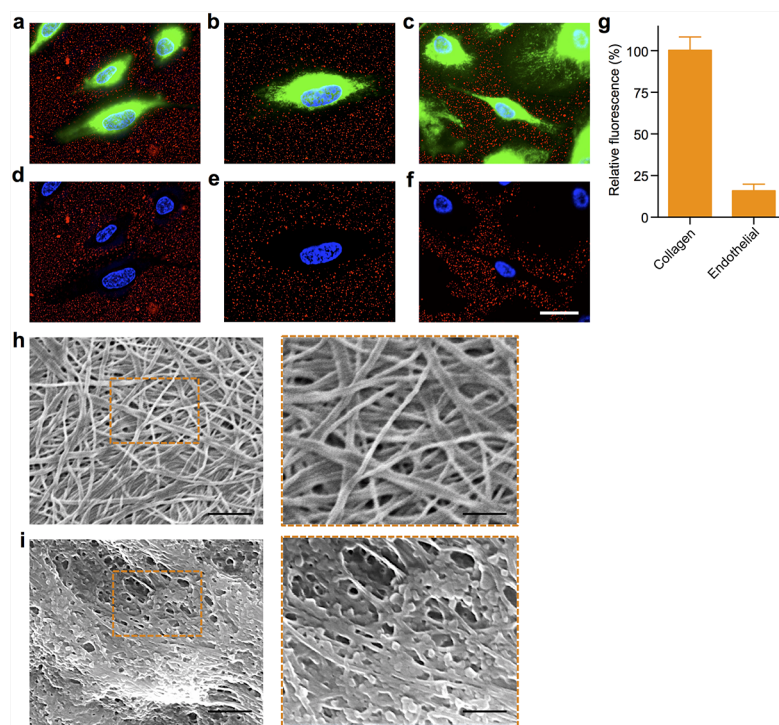


Extended Data Figure 4 | Platelet membrane sidedness on PNPs.

a, Transmission electron micrograph of PNPs primary-stained with anti-CD47 (intracellular), secondary-stained with immunogold, and negatively stained with 2% vanadium. The immunogold staining revealed presence of intracellular CD47 domains on collapsed platelet membrane vesicles, but not on PNPs.

b, Transmission electron micrograph of PNPs primary-stained with anti-CD47 (extracellular), secondary-stained with immunogold, and negatively stained with 2% vanadium. PNPs were shown to display extracellular CD47 domains.

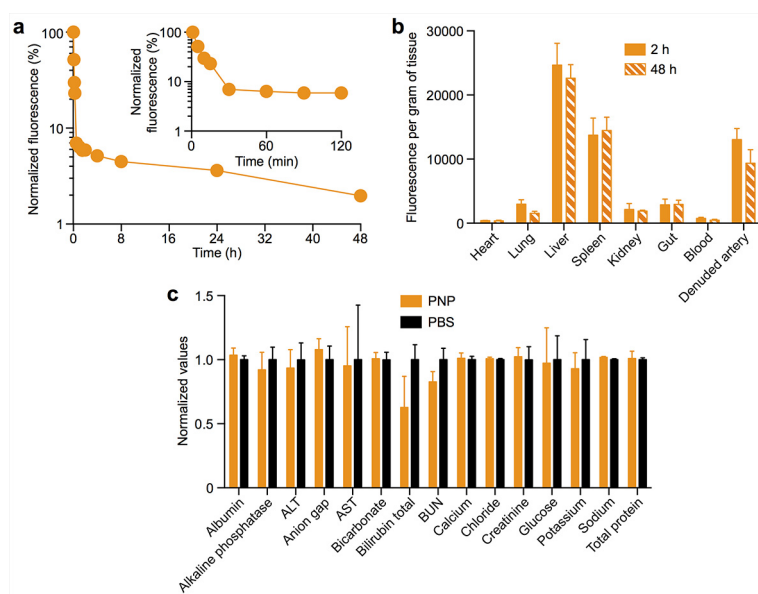
All scale bars, 100 nm. **c**, 2 μ m polystyrene beads were functionalized with anti-CD47 against the protein's extracellular domain, anti-CD47 against the protein's intracellular domain, or a sham antibody. Flow cytometric analysis of the different beads after DiD-loaded PNP incubation showed the highest particle retention to beads functionalized with anti-CD47 against the protein's extracellular domain. **d**, Normalized fluorescence intensity of PNP retention to the different antibody-functionalized beads. Bars represent means \pm s.e.m.



Extended Data Figure 5 | PNP binding to collagen and extracellular matrix.

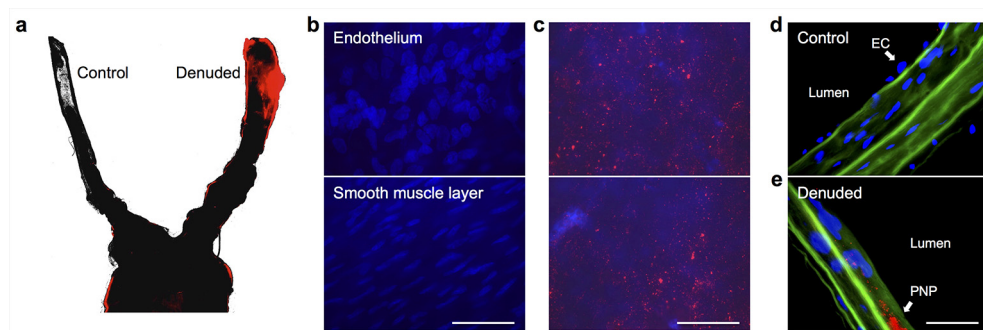
a–f, Collagen-coated tissue culture slides seeded with human umbilical vein endothelial cells (HUVECs) were incubated with PNP solution for 30 s. Fluorescence microscopy samples demonstrate differential PNP adherence to exposed collagen versus covered endothelial surfaces. **a–c**, Representative fluorescence images visualizing DiD-loaded PNPs (red), cellular cytosol (green), and cellular nuclei (blue). **d–f**, Images showing only the red and blue channels to highlight the differential localization of PNPs. Scale bar, 10 μ m.

g, Fluorescence quantification of PNP per unit area on collagen and endothelial surfaces. Bars represent means \pm s.d. ($n = 10$). **h, i**, PNP adherence to arterial extracellular matrix (ECM) as visualized by SEM. **h**, SEM images of the ECM of a decellularized human umbilical cord artery. Left, scale bar, 1 μ m; right, scale bar, 500 nm. **i**, SEM images of the ECM of a decellularized human umbilical cord artery after PNP incubation. Left, scale bar, 1 μ m; right, scale bar, 500 nm.



Extended Data Figure 6 | Pharmacokinetics, biodistribution and safety of PNPs. **a**, DiD-loaded PNPs were injected intravenously through the tail vein of Sprague–Dawley rats. At various time points, blood was withdrawn via tail vein blood sampling for fluorescence quantification to evaluate the systemic circulation lifetime of the nanoparticles ($n = 6$). **b**, Biodistribution of the PNP nanoparticles in balloon-denuded Sprague–Dawley rats at 2 h and 48 h after intravenous nanoparticle administration through the tail vein ($n = 6$).

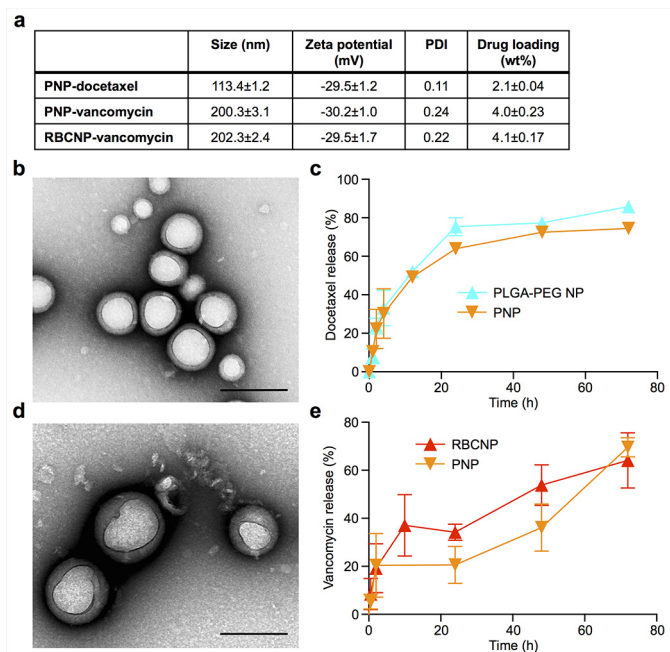
c, Comprehensive metabolic panel of rats after injections with human-derived PNPs and PBS ($n = 6$). The rats received intravenous injections of PNPs and PBS on day 0 and day 5, and the blood test conducted on day 10 did not reveal significant changes between the two groups, indicating normal liver and kidney functions after the PNP administration. All bars and markers represent means \pm s.d.



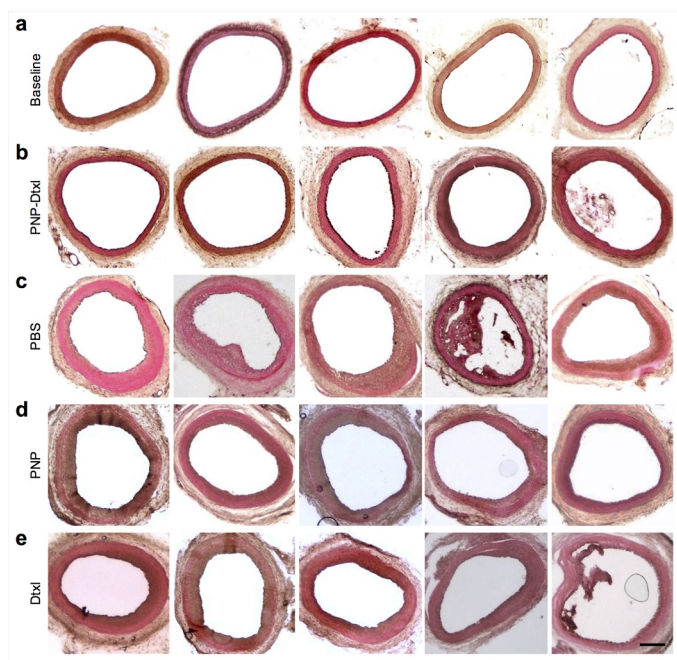
Extended Data Figure 7 | PNP targeting of damaged vasculatures upon intravenous injection to rats with angioplasty-induced arterial denudation.

a, Fluorescence microscopy of the aortic branch revealed selective PNP binding to the denuded artery (right) as opposed to the undamaged artery (left) (PNP fluorescence in red). **b**, Fluorescence images acquired from the control artery, which did not reveal PNP fluorescence upon focusing on either the endothelium (top) or the smooth muscle layer (bottom) (nuclei in blue). **c**, Fluorescence images acquired from the denuded artery, which revealed

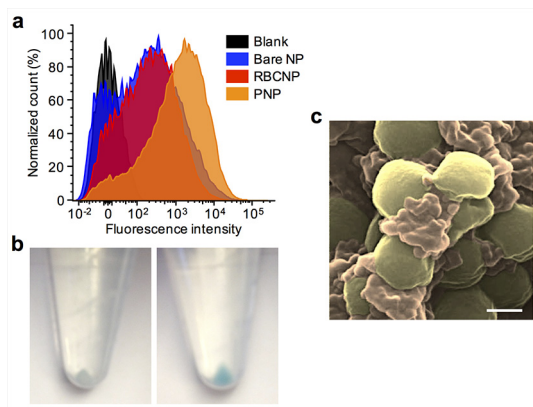
significant PNP retention as fluorescent punctates (PNP fluorescence in red) above the smooth muscle layer. **d**, Fluorescence image of arterial cross-section acquired from the control artery, which showed nuclei of endothelial cells above the collagen layer (autofluorescence in green) and an absence of PNP fluorescence. **e**, Fluorescence image of arterial cross-section acquired from the denuded artery, which showed PNP retention as fluorescent punctates on the collagen layer (PNP fluorescence in red; collagen autofluorescence in green) and an absence of endothelial cell nuclei. All scale bars, 100 μm .



Extended Data Figure 8 | Characterizations of drug-loaded cell membrane cloaked nanoparticles. **a**, Physicochemical properties of drug-loaded cell membrane cloaked nanoparticles. **b**, TEM visualization of docetaxel-loaded PNPs (PNP-Dtxl). Scale bar, 200 nm. **c**, Drug release profile of PNP-Dtxl compared to polyethylene glycol (PEG)-PLGA diblock nanoparticles of equivalent size and docetaxel loading ($n = 3$). **d**, TEM visualization of vancomycin-loaded PNPs (PNP-Vanc). Scale bar, 200 nm. **e**, Drug release profiles of PNP-Vanc and RBCNP-Vanc ($n = 3$). Bars represent means \pm s.d.



Extended Data Figure 9 | Treatment of an experimental rat model of coronary restenosis. a–e, H&E-stained arterial cross-sections reveal the vascular structure of non-damaged arteries (serving as baseline, a) and denuded arteries after treatment with PNP-Dtxl (b), PBS (c), PNP with no docetaxel content (d), or free docetaxel (e). Scale bar, 200 μ m.



Extended Data Figure 10 | PNP adherence to MRSA252 bacteria. **a**, Flow cytometric analysis of MRSA252 bacteria after incubation with different DiD-loaded nanoformulations. **b**, Pellets of MRSA252 after incubation with DiD-loaded RBCNPs (left) and DiD-loaded PNPs (right) show differential retention of nanoformulation with MRSA252 upon pelleting of the bacteria. **c**, A pseudocoloured SEM image of PNPs binding to MRSA252 under high magnification (MRSA coloured in gold, PNP coloured in orange). Scale bar, 400 nm.

The soft palate is an important site of adaptation for transmissible influenza viruses

Seema S. Lakdawala^{1†}, Akila Jayaraman², Rebecca A. Halpin³, Elaine W. Lamirande¹, Angela R. Shih¹, Timothy B. Stockwell³, Xudong Lin³, Ari Simenauer³, Christopher T. Hanson¹, Leatrice Vogel¹, Myeisha Paskel¹, Mahnaz Minai⁴, Ian Moore⁴, Marlene Orandle^{4†}, Suman R. Das³, David E. Wentworth^{3†}, Ram Sasisekharan² & Kanta Subbarao¹

Influenza A viruses pose a major public health threat by causing seasonal epidemics and sporadic pandemics. Their epidemiological success relies on airborne transmission from person to person; however, the viral properties governing airborne transmission of influenza A viruses are complex. Influenza A virus infection is mediated via binding of the viral haemagglutinin (HA) to terminally attached $\alpha 2,3$ or $\alpha 2,6$ sialic acids on cell surface glycoproteins. Human influenza A viruses preferentially bind $\alpha 2,6$ -linked sialic acids whereas avian influenza A viruses bind $\alpha 2,3$ -linked sialic acids on complex glycans on airway epithelial cells^{1,2}. Historically, influenza A viruses with preferential association with $\alpha 2,3$ -linked sialic acids have not been transmitted efficiently by the airborne route in ferrets^{3,4}. Here we observe efficient airborne transmission of a 2009 pandemic H1N1 (H1N1pdm) virus (A/California/07/2009) engineered to preferentially bind $\alpha 2,3$ -linked sialic acids. Airborne transmission was associated with rapid selection of virus with a change at a single HA site that conferred binding to long-chain $\alpha 2,6$ -linked sialic acids, without loss of $\alpha 2,3$ -linked sialic acid binding. The transmissible virus emerged in experimentally infected ferrets within 24 hours after infection and was remarkably enriched in the soft palate, where long-chain $\alpha 2,6$ -linked sialic acids predominate on the nasopharyngeal surface. Notably, presence of long-chain $\alpha 2,6$ -linked sialic acids is conserved in ferret, pig and human soft palate. Using a loss-of-function approach with this one virus, we demonstrate that the ferret soft palate, a tissue not normally sampled in animal models of influenza, rapidly selects for transmissible influenza A viruses with human receptor ($\alpha 2,6$ -linked sialic acids) preference.

Receptor-binding specificity is an important determinant of host-range restriction and transmission of influenza A viruses (refs 4, 5 and reviewed in ref. 6). The ability of zoonotic influenza A viruses to transmit via the airborne route increases their pandemic potential⁷. Recently, several investigators have attempted to identify viral determinants of airborne transmission by generating transmissible H5 and H7 avian influenza A viruses^{8–10}. We approached the question differently and used an epidemiologically successful influenza A virus in which we altered receptor preference from the human ($\alpha 2,6$ -linked sialic acids) to the avian receptor ($\alpha 2,3$ -linked sialic acids).

We previously generated H1N1pdm virus variants with highly specific binding to either $\alpha 2,6$ -linked or $\alpha 2,3$ -linked sialic acids (referred to as $\alpha 2,6$ or $\alpha 2,3$ H1N1pdm virus, respectively)¹¹. The $\alpha 2,3$ H1N1pdm virus was generated by introducing four amino acid mutations in the receptor binding site of HA (D187E, I216A, D222G and E224A)¹¹. Unexpectedly, the $\alpha 2,6$ and $\alpha 2,3$ H1N1pdm viruses transmitted via the airborne route equally well in ferrets (Fig. 1 and Supplementary Table 1) and with a similar efficiency as observed previously for wild-type H1N1pdm virus^{12–15}.

A delay in peak viral shedding was noted in the airborne-contact animals (see Fig. 1 legend for details) in the $\alpha 2,3$ virus group (red arrows, Fig. 1), suggesting that the virus evolves before transmission. Deep sequence analysis of viral RNA (vRNA) extracted from nasal washes of $\alpha 2,3$ H1N1pdm virus-infected ferrets revealed a mixed population at amino acid position 222 (H1 numbering) with the engineered glycine (G) and wild-type aspartic acid (D), while the other three engineered changes in the HA were retained (Fig. 2a and Supplementary Table 2). Interestingly, the vRNA from the nasal washes of airborne-contact ferrets contained only the G222D HA mutation (Fig. 2a and Supplementary Table 2), suggesting that this sequence at amino acid 222 in the $\alpha 2,3$ H1N1pdm virus was associated with airborne transmission. The virus inoculum did not contain a mixture at this residue (Fig. 2a), and associated changes were not observed in the neuraminidase gene (Supplementary Table 3).

A D222G change in the 2009 H1N1pdm virus HA has occurred in natural isolates, and reports suggest an association with increased virulence in humans and no effect on airborne transmission^{16–18}. Theoretical structural analysis suggests that the G222D reversion makes the receptor binding site better suited to bind $\alpha 2,6$ -linked sialic acids while retaining contacts with $\alpha 2,3$ -linked sialic acids via glutamic acid at amino acid 187 (Extended Data Fig. 1). Glycan binding data corroborated this structural prediction because the G222D mutation caused no change in $\alpha 2,3$ -linked sialic acid binding but substantially increased binding to long-chain $\alpha 2,6$ -linked sialic acids (Fig. 2b).

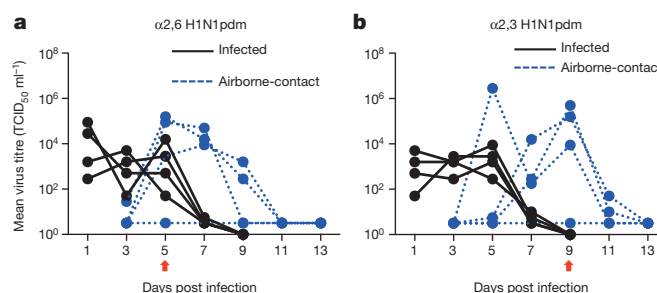


Figure 1 | Airborne transmission of receptor-specific H1N1pdm viruses. Transmission studies were performed with 4 pairs of animals (8 animals total) in double secure cages with perforated dividers¹². One ferret in each pair was infected with 10^6 50% tissue culture infectious dose (TCID₅₀) of the indicated virus; a naive ferret (referred to as airborne-contact) was introduced into the adjacent compartment 24 h later. Nasal secretions were collected every other day for 14 days. Viral titres from the nasal secretions are graphed for each infected or airborne-contact animal. Transmission of $\alpha 2,6$ H1N1pdm (a) and $\alpha 2,3$ H1N1pdm (b) viruses was similar. The red arrow indicates the peak day of viral shedding for airborne contact animals.

¹Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. ²Department of Biological Engineering, Koch Institute for Integrative Cancer Research, Singapore-MIT Alliance for Research and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ³J. Craig Venter Institute, Rockville, Maryland 20850, USA. ⁴Comparative Medicine Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [†]Present addresses: Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15217, USA (S.S.L.); National Institute of Occupational Safety and Health, Centers for Disease Control and Prevention, Morgantown, West Virginia, 26505 USA (M.O.); Virology Surveillance and Diagnosis Branch, Influenza Division, Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA (D.E.W.).

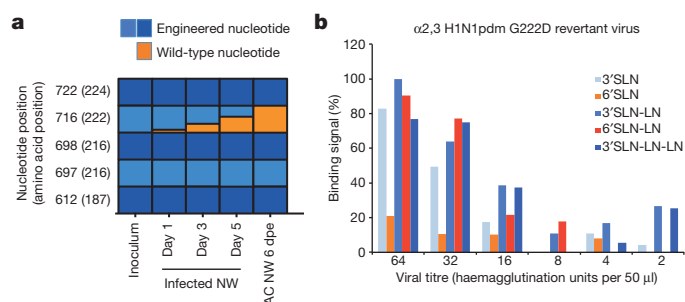


Figure 2 | Characterization of transmissible $\alpha 2,3$ H1N1pdm viruses. **a**, Deep sequencing of the $\alpha 2,3$ H1N1pdm inoculum, nasal wash (NW) from an infected ferret on 1, 3 and 5 dpi, and nasal wash from one airborne-contact (AC) animal on 6 days post-exposure (dpe) revealed a reversion at residue 222 from G to D. This data is representative of the three transmission pairs that resulted in infection of the airborne-contact animals. Graphical representation of the proportion of reads at each engineered nucleotide is shown. Blue shading represents the $\alpha 2,3$ -linked sialic acid engineered nucleotide and orange represents the wild-type nucleotide residue. All other engineered nucleotides were maintained. **b**, A G222D reversion, in the context of the other engineered mutations, affects the glycan specificity of the $\alpha 2,3$ H1N1pdm virus. The glycans are indicated in the key to the figure and are defined in the Methods; orange colours represent $\alpha 2,6$ -linked sialic acids and blue colours represent $\alpha 2,3$ -linked sialic acids. H1 numbering is used for all amino acid positions.

Previous reports have demonstrated the importance of $\alpha 2,6$ -linked sialic acid binding for transmission^{4,5,19}. We now demonstrate conclusively that airborne transmission requires gain of long-chain $\alpha 2,6$ -linked sialic acid binding and, contrary to previous suggestions⁴, loss of $\alpha 2,3$ -linked sialic acid binding is not necessary.

The presence of a distinct and identifiable HA sequence in the transmissible virus allowed us to determine whether it emerges in a specific area of the respiratory tract of experimentally infected ferrets. Tissue sections and samples from the upper and lower respiratory tract were collected several days post-infection (dpi) from groups of 3 ferrets infected with the $\alpha 2,3$ H1N1pdm virus. Virus was detected in all ferrets and all samples (Extended Data Fig. 2). Deep sequencing of vRNA from both the upper and lower respiratory tract revealed a mixed population at residue 222 (Fig. 3). Surprisingly, vRNA from the soft palate was remarkably and uniquely enriched for the G222D virus on 1 dpi and $\geq 90\%$ of the sequences encoded D222 at 3 dpi (Fig. 3c). All other engineered mutations were maintained (Extended Data Fig. 3). These data suggest that the G222D revertant virus was actively selected in the ferret soft palate.

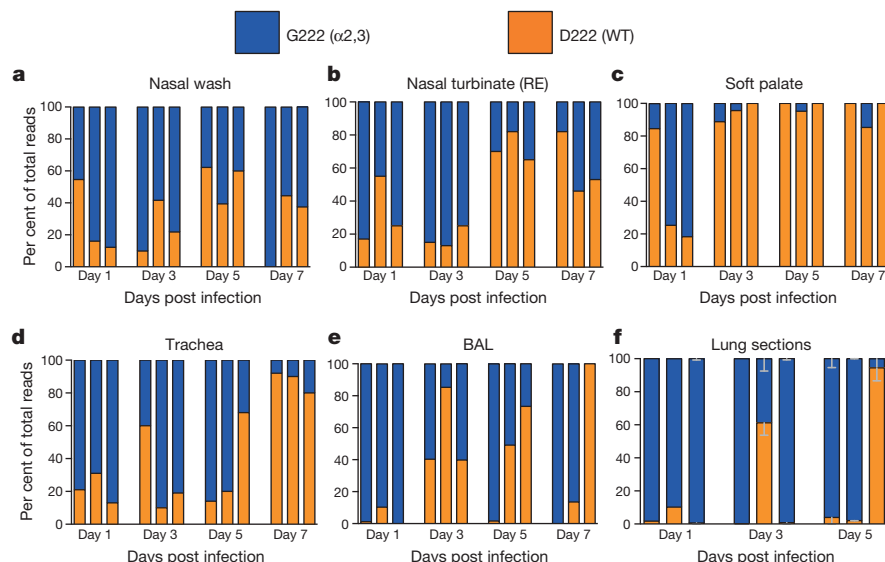


Figure 3 | Emergence of the $\alpha 2,3$ G222D H1N1pdm virus in the ferret respiratory tract. **a–f**, Different samples from the ferret respiratory tract: nasal wash (**a**), respiratory epithelium (RE) of nasal turbinates (**b**), soft palate (**c**), trachea (**d**), bronchoalveolar lavage (BAL) (**e**), and combined right, middle and left cranial lung sections (**f**) were collected from three animals each on 1, 3, 5 and 7 dpi. The respiratory epithelium region of the nasal turbinates is depicted in Extended Data Fig. 6h. The HA gene from virus populations in these samples were deep sequenced and the proportion of reads with D at position 222 is shown in orange, and G is shown in blue. Each bar represents a single animal. The standard error between the right and left lung sections is shown in **f**.

To determine whether the rapid enrichment of G222D revertant virus in the soft palate was responsible for infection of the airborne-contact animal, we performed an airborne transmission study where naive ferrets were exposed to experimentally infected donor ferrets for only 2 days. Surprisingly, even within this shortened exposure time, two airborne-contact animals shed virus and 3 out of 4 airborne-contact animals seroconverted (Extended Data Fig. 4 and Supplementary Table 1). Sequence analysis of vRNA from the two airborne-contact animals with detectable virus in the nasal washes revealed presence of the G222D revertant. These data suggest that the selection of the $\alpha 2,3$ H1N1pdm virus with the D222 sequence occurs within 3 dpi in the donor ferret and that the airborne-contact ferrets were possibly infected with virus originating in the soft palate because there was nearly complete selection of the G222D mutant by 3 dpi in this tissue.

The soft palate, with mucosal surfaces facing the oral cavity and nasopharynx, is not usually examined in animal models of influenza. To understand what drives the enrichment of the long-chain $\alpha 2,6$ -linked-sialic acid-binding G222D revertant virus at this site, we stained the soft palate with lectins specific for $\alpha 2,6$ or $\alpha 2,3$ sialic acids (Extended Data Fig. 5). The ciliated respiratory epithelium and mucus secreting goblet cells in the respiratory epithelium and submucosal glands (SMG) contained $\alpha 2,6$ -linked sialic acids (SNA staining) (Extended Data Fig. 5). Expression of $\alpha 2,3$ -linked sialic acid (MAL II staining) was present in the connective tissue underlying the respiratory epithelium and in the serous cells of the SMG. Using a purified HA protein (SC18) that selectively binds long-chain $\alpha 2,6$ -linked sialic acids²⁰, we found high expression of long-chain $\alpha 2,6$ -linked sialic acids in the soft palate compared to the trachea and lungs of ferrets (Fig. 4 and Extended Data Fig. 6). A recent report detailing the glycan profile of the ferret respiratory tract confirms that the soft palate abundantly expresses $\alpha 2,6$ sialylated LacNAc structures²¹, similar to the long-chain $\alpha 2,6$ -linked sialic acids recognized by SC18 HA. Interestingly, both the respiratory epithelium and olfactory epithelium from the nasal turbinates of ferrets expressed high levels of long-chain $\alpha 2,6$ -linked sialic acid, but the respiratory epithelium of the nasal turbinates was not enriched for the G222D mutant (Fig. 3b and Extended Data Fig. 6). These data suggest that the soft palate is unusual in driving selection for the G222D virus.

To determine the relevance for humans, we evaluated the expression of long-chain $\alpha 2,6$ -linked sialic acids in the soft palate of humans and pigs. Interestingly, expression of long-chain $\alpha 2,6$ -linked sialic acids was conserved on the respiratory epithelium and goblet cells of the soft palate of both species (Fig. 4). In addition, staining with plant lectins specific for $\alpha 2,6$ -linked or $\alpha 2,3$ -linked sialic acids (Extended Data Fig. 7) revealed

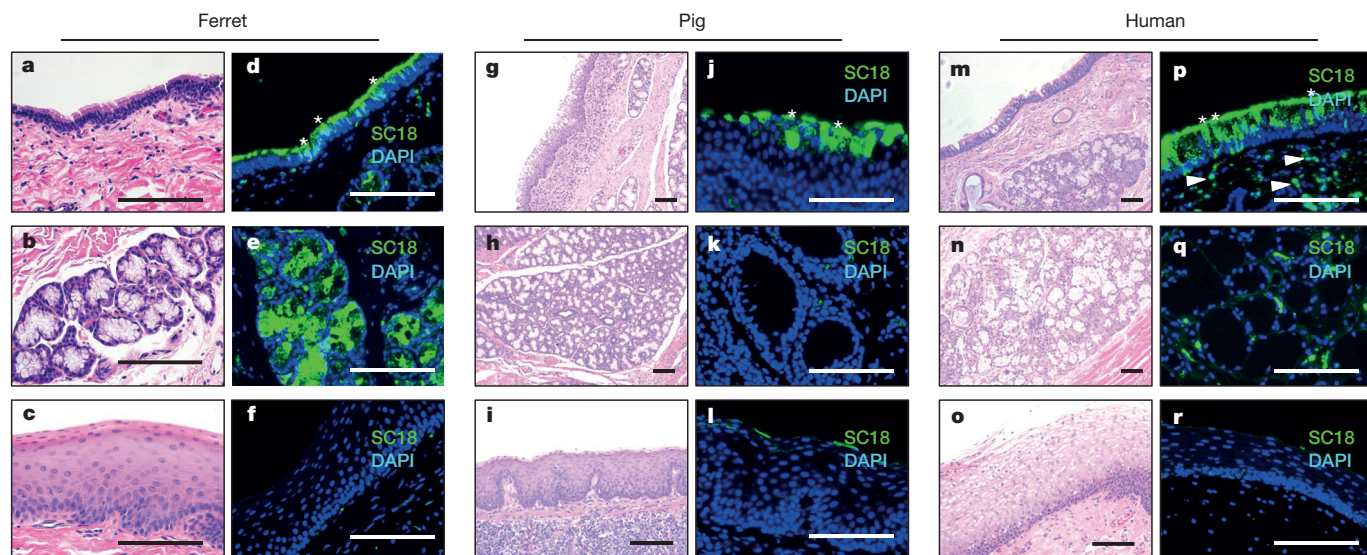


Figure 4 | Comparison of long-chain α 2,6-linked sialic acid expression in the soft palate of ferrets, pigs and humans. a–c, g–i, m–o, Haematoxylin and eosin staining of the soft palate from an uninfected ferret (a–c), pig (g–i) and human (m–o) highlights the nasopharyngeal, SMG and oral surfaces. d–f, j–l, p–r, Purified SC18 HA was used to define long-chain α 2,6-linked sialic acids in these sections from an uninfected ferret (d–f), pig (j–l) and human (p–r). Staining of the nasopharyngeal surface is depicted for each species across

the first row, SMG in the second row and oral surface on the last row. At least two independent tissue samples were stained and analysed for each species. A sialidase-A-treated control was run for each sample to ensure specificity of SC18 HA (not shown). Scale bars, 100 μ m in all images. Asterisks highlight SC18-positive goblet cells and white arrowheads indicate SC18-positive plasma cells in human soft palate.

that α 2,6-linked sialic acids were present on the nasopharyngeal surface and SMG of both pigs and humans. Expression of α 2,3-linked sialic acids was detected in the basal cells of the oral surface and on the nasopharyngeal surface of the human soft palate; these findings are consistent with reports describing the sialic acid distribution in the human nasopharynx²². Other investigators have also reported replication of seasonal and pandemic influenza A viruses in tissue sections obtained from the human nasopharynx²³. Taken together, these data highlight the importance of the nasopharynx, of which the soft palate forms the floor, as a site for host adaptation of influenza A viruses.

Influenza A virus infection of the soft palate may contribute to airborne transmission by providing a mucin-rich microenvironment for generation of airborne virus during coughing, sneezing or breathing. Infection with α 2,3 H1N1pdm virus resulted in severe inflammation and necrosis of the respiratory epithelial cells and SMG in the soft palate (Extended Data Fig. 8). Since the soft palate is innervated by the trigeminal nerve, inflammation of this tissue could stimulate sneezing. Alternatively, the soft palate may be the site where infection is initiated during airborne transmission; therefore binding to this tissue would provide a fitness advantage.

These results, albeit with one virus, enhance our understanding of the properties necessary for airborne transmission of influenza A viruses in the ferret model. Loss of α 2,3-linked sialic acid specificity is not necessary but gain of long-chain α 2,6-linked sialic acid binding is critical for efficient airborne transmission of influenza A viruses. H7N9 viruses from China show dual receptor binding but variable airborne transmission efficiency in ferrets^{24,25}. Interestingly, the 1918 H1N1 virus (A/New York/1/18), which has a similar sialic acid binding preference as the α 2,3 H1N1pdm virus, did not transmit via the airborne route or adapt within the ferret host⁴, suggesting that the 2009 H1N1pdm virus may be unusual for this rapid adaptation. However, the detection of a mutation that enhanced α 2,6-linked sialic acid binding in nasal washes of ferrets infected with avian H2 viruses was recently reported²⁶, demonstrating that rapid adaptation of influenza A viruses to gain human receptor preference occurs in other influenza A virus subtypes as well.

Studies with transmissible H5 viruses suggest that increased pH and thermal stability of the HA enhance airborne transmission^{8,9,27}.

Although we did not observe adaptive mutations in the HA stalk of the α 2,3 H1N1pdm virus, perhaps because H1N1pdm HA is already adapted to humans, a mixed population was observed at four lysine residues around the receptor binding site (Extended Data Fig. 9 and Supplementary Table 2). Some are known to be egg adaptive mutations²⁸ or are components of the proposed positively charged 'lysine fence' around the base of the receptor binding site, positioned to anchor the *N*-acetylneuraminic acid and galactose sugar of α 2,3-linked and α 2,6-linked sialic acid glycans²⁹. Interestingly, the lysine residues were restored in the vRNA isolated from nasal washes of airborne-contact ferrets and the soft palate of experimentally infected ferrets (Extended Data Figs 9 and 10).

Taken together with our previously published data, long-chain α 2,6-linked sialic acid binding and a highly active neuraminidase contribute to the airborne transmission of the H1N1pdm virus^{12,30}. Importantly, we have identified the previously overlooked soft palate as an important site of isolation of transmissible virus and perhaps the initial site of infection. Analysis of the replicative fitness of influenza A viruses in this tissue may be warranted in assessment of their pandemic potential.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 October 2014; accepted 5 August 2015.

Published online 23 September 2015.

- Shinya, K. *et al.* Avian flu: influenza virus receptors in the human airway. *Nature* **440**, 435–436 (2006).
- van Riel, D. *et al.* H5N1 virus attachment to lower respiratory tract. *Science* **312**, 399 (2006).
- Maines, T. R. *et al.* Lack of transmission of H5N1 avian-human reassortant influenza viruses in a ferret model. *Proc. Natl Acad. Sci. USA* **103**, 12121–12126 (2006).
- Tumpey, T. M. *et al.* A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission. *Science* **315**, 655–659 (2007).
- Pappas, C. *et al.* Receptor specificity and transmission of H2N2 subtype viruses isolated from the pandemic of 1957. *PLoS ONE* **5**, e11158 (2010).
- Cauldwell, A. V., Long, J. S., Moncorge, O. & Barclay, W. S. Viral determinants of influenza A virus host range. *J. Gen. Virol.* **95**, 1193–1210 (2014).
- Lakdawala, S. S. & Subbarao, K. The ongoing battle against Influenza: The challenge of flu transmission. *Nature Med.* **18**, 1468–1470 (2012).

8. Herfst, S. *et al.* Airborne transmission of influenza A/H5N1 virus between ferrets. *Science* **336**, 1534–1541 (2012).
9. Imai, M. *et al.* Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* **486**, 420–428 (2012).
10. Sutton, T. C. *et al.* Airborne transmission of highly pathogenic H7N1 influenza virus in ferrets. *J. Virol.* **88**, 6623–6635 (2014).
11. Lakdawala, S. S. *et al.* Receptor specificity does not affect replication or virulence of the 2009 pandemic H1N1 influenza virus in mice and ferrets. *Virology* **446**, 349–356 (2013).
12. Lakdawala, S. S. *et al.* Eurasian-origin gene segments contribute to the transmissibility, aerosol release, and morphology of the 2009 pandemic H1N1 influenza virus. *PLoS Pathog.* **7**, e1002443 (2011).
13. Itoh, Y. *et al.* *In vitro* and *in vivo* characterization of new swine-origin H1N1 influenza viruses. *Nature* **460**, 1021–1025 (2009).
14. Maines, T. R. *et al.* Transmission and pathogenesis of swine-origin 2009 A(H1N1) influenza viruses in ferrets and mice. *Science* **325**, 484–487 (2009).
15. Munster, V. J. *et al.* Pathogenesis and transmission of swine-origin 2009 A(H1N1) influenza virus in ferrets. *Science* **325**, 481–483 (2009).
16. Liu, Y. *et al.* Altered receptor specificity and cell tropism of D222G hemagglutinin mutants isolated from fatal cases of pandemic A(H1N1) 2009 influenza virus. *J. Virol.* **84**, 12069–12074 (2010).
17. Mak, G. C. *et al.* Association of D222G substitution in haemagglutinin of 2009 pandemic influenza A(H1N1) with severe disease. *Euro Surveill.* **15**, 19534 (2010).
18. Belser, J. A. *et al.* Effect of D222G mutation in the hemagglutinin protein on receptor binding, pathogenesis and transmissibility of the 2009 pandemic H1N1 influenza virus. *PLoS ONE* **6**, e25091 (2011).
19. Zhang, Y. *et al.* Key molecular factors in hemagglutinin and PB2 contribute to efficient transmission of the 2009 H1N1 pandemic influenza virus. *J. Virol.* **86**, 9666–9674 (2012).
20. Srinivasan, A. *et al.* Quantitative biochemical rationale for differences in transmissibility of 1918 pandemic influenza A viruses. *Proc. Natl Acad. Sci. USA* **105**, 2800–2805 (2008).
21. Jia, N. *et al.* Glycomic characterisation of respiratory tract tissues of ferrets: implications for its use in influenza virus infection studies. *J. Biol. Chem.* **289**, 28489–28504 (2014).
22. Nicholls, J. M., Bourne, A. J., Chen, H., Guan, Y. & Peiris, J. S. Sialic acid receptor detection in the human respiratory tract: evidence for widespread distribution of potential binding sites for human and avian influenza viruses. *Respir. Res.* **8**, 73 (2007).
23. Chan, R. W., Chan, M. C., Nicholls, J. M. & Malik Peiris, J. S. Use of *ex vivo* and *in vitro* cultures of the human respiratory tract to study the tropism and host responses of highly pathogenic avian influenza A (H5N1) and other influenza viruses. *Virus Res.* **178**, 133–145 (2013).
24. Belser, J. A. *et al.* Pathogenesis and transmission of avian influenza A (H7N9) virus in ferrets and mice. *Nature* **501**, 556–559 (2013).
25. Richard, M. *et al.* Limited airborne transmission of H7N9 influenza A virus between ferrets. *Nature* **501**, 560–563 (2013).
26. Pappas, C. *et al.* Assessment of transmission, pathogenesis and adaptation of H2 subtype influenza viruses in ferrets. *Virology* **477**, 61–71 (2015).
27. Linster, M. *et al.* Identification, characterization, and natural selection of mutations driving airborne transmission of A/H5N1 virus. *Cell* **157**, 329–339 (2014).
28. Chen, Z. *et al.* Generation of live attenuated novel influenza virus A/California/7/09 (H1N1) vaccines with high yield in embryonated chicken eggs. *J. Virol.* **84**, 44–51 (2010).
29. Soundararajan, V. *et al.* Extrapolating from sequence—the 2009 H1N1 ‘swine’ influenza virus. *Nature Biotechnol.* **27**, 510–513 (2009).
30. Yen, H. L. *et al.* Hemagglutinin-neuraminidase balance confers respiratory-droplet transmissibility of the pandemic H1N1 influenza virus in ferrets. *Proc. Natl Acad. Sci. USA* **108**, 14264–14269 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was supported in part by the Intramural Research Program of NIAID, NIH and with federal funds from NIAID, NIH, DHHS under contract number HHSN272200900007C and NIAID/NIH Genomic Centers for Infectious Diseases (GCID) program (U19-AI-110819). This manuscript was reviewed by the NIH's Intramural Research Program's Committee on Dual Use Research of Concern (DURC), who concluded that the methods and results do not meet DURC criteria. We thank the NIAID Comparative Medical Branch for technical assistance, Subbarao laboratory members for critical input, N. B. Fedorova from JCVI for technical help, X. J. Meng (Virginia Tech College of Veterinary Medicine) and P. Pineyro (Iowa State University) for pig soft palate tissues, and the Consortium for Functional Glycomics for providing glycans for the glycan array analysis. The data for this manuscript and its preparation were generated while D.E.W. was employed at JCVI. The opinions expressed in this article are the authors' own and do not reflect the views of the Centers for Disease Control, the Department of Health and Human Services, or the United States government. A.J. and R.S. are supported in part by NIH Merit Award (R37 GM057073-13), National Research Foundation supported Interdisciplinary Research group in Infectious Diseases of SMART (Singapore MIT alliance for Research and Technology) and the Skolkovo Foundation supported Infectious Diseases Center at MIT.

Author Contributions S.S.L., A.J., R.S., D.E.W. and K.S. designed the study. S.S.L., A.J., E.W.L., A.R.S., X. L., A.S., C.T.H., L.V., M.P. and M.M. performed the experiments. S.S.L., R.A.H., T.B.S., I.M., M.O. and S.R.D. analysed the data. S.S.L., K.S., A.J. and R.S. wrote the paper.

Author Information The sequences detailed in this manuscript can be found in GenBank under accession numbers CY184674–CY185309. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.S. (rams@mit.edu) and K.S. (ksubbarao@niaid.nih.gov).

METHODS

Ethics statement and animal studies. This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The National Institutes of Health Animal Care and Use Committee (ACUC) approved the animal experiments that were conducted. All studies were conducted under ABSL2 conditions and all efforts were made to minimize suffering. No statistical methods were used to predetermine sample size. In our animal study protocol, we state that the number of animals in each experimental group varies, and is based on our prior experience. We use the minimum number of animals per group that will provide meaningful results. Randomization was not used to allocate animals to experimental groups and the animal studies were not blinded.

Virus rescue. The 2009 H1N1pdm virus used in this study is influenza A/California/07/2009. Generation and characterization of the α 2,3 H1N1pdm and α 2,6 H1N1pdm viruses have been described previously¹¹. Genomic sequencing and dose-dependent glycan binding assays confirmed the identity and receptor specificity of viruses generated by reverse genetics. All experiments were performed using viruses passaged no more than three times in MDCK cells (ATCC) or 10-day old embryonated chicken eggs.

Ferret transmission study. All ferrets (*Mustela putorius furo*) were obtained from Triple F Farms (Sayre, PA) and screened by haemagglutination inhibition (HAI) assay before infection to ensure that they were naive to seasonal influenza A and B viruses and the viruses used in this study. The transmission studies were conducted in adult ferrets as previously described¹², male and female ferrets were used in a 3:1 ratio and sample size was based on the capacity of the transmission cages. Ferrets reaching 15–20% weight loss were provided with enriched diet and monitored closely by veterinary staff for altered behaviour.

Environmental conditions inside the laboratory were monitored daily and were consistently $19 \pm 1^\circ\text{C}$ and $56 \pm 2\%$ relative humidity. The transmission experiments were conducted in the same room, to minimize any effects of caging and airflow differences on aerobiology. On day 0 four animals were infected intranasally with 10^6 TCID₅₀ of either α 2,3 H1N1pdm or α 2,6 H1N1pdm virus and placed into the transmission cage. Twenty-four hours post-infection, a naive animal (airborne-contact) was placed into the transmission cage on the other side of a perforated stainless steel barrier. The airborne-contact ferrets were always handled before the infected ferrets. Nasal washes were collected and clinical signs were recorded on alternate days from days 0 to 14. Great care was taken during nasal wash collections and husbandry to ensure that direct contact did not occur between the ferrets. On 14 days post-infection (dpi), blood was collected from each animal for serology. The shortened exposure time study was done similarly except 48 h after the naive recipient animal (airborne contact) was placed into the transmission cage the ferrets were separated into micro-isolator cages. Infected ferrets in the shortened exposure time study were killed on 7 dpi and the airborne contact animals were killed on 21 dpi. The airborne-contact animals were always handled before infected ferrets and all husbandry tools were decontaminated three times between handling of each airborne-contact animal.

Dose-dependent direct binding of influenza viruses. To determine the receptor specificity of the G222D α 2,3 H1N1pdm virus, virus from the nasal wash of a single airborne-contact animal on day 6 post-exposure was propagated once in MDCK cells. This virus stock was inactivated with betapropiolactone and the haemagglutination titre was determined. For the glycan binding assay, 50 μl of 2.4 μM biotinylated glycans were added to wells of streptavidin-coated high binding capacity 384-well plates (Pierce) and incubated overnight at 4°C . The glycans included were 3'SLN, 3'SLN-LN, 3'SLN-LN-LN, 6'SLN and 6'SLN-LN (LN corresponds to lactosamine (Gal β 1-4GlcNAc) and 3'SLN and 6'SLN respectively correspond to Neu5Ac α 2-3 and Neu5Ac α 2-6 linked to LN) that were obtained from the Consortium of Functional Glycomics (<http://www.functionalglycomics.org>). The inactivated G222D virus was diluted to 250 μl with $1 \times$ PBS + 1% BSA. 50 μl of diluted virus was added to each of the glycan-coated wells and incubated overnight at 4°C . This was followed by three washes with $1 \times$ PBST (1X PBS + 0.1% Tween-20) and three washes with $1 \times$ PBS. The wells were blocked with $1 \times$ PBS + 1% BSA for 2 h at 4°C followed by incubation with primary antibody (ferret anti-CA07/09 antisera; 1:200 diluted in $1 \times$ PBS + 1% BSA) for 5 h at 4°C . This was followed by three washes with $1 \times$ PBST and three washes with $1 \times$ PBS. Finally, the wells were incubated with the secondary antibody (goat anti-ferret HRP conjugated antibody from Rockland; 1:200 diluted in 1X PBS + 1% BSA). The wells were washed with $1 \times$ PBST and $1 \times$ PBS as before. The binding signals were determined based on the HRP activity using the Amplex Red Peroxidase Assay (Invitrogen) according to the manufacturer's instructions. Negative controls were uncoated wells (without any glycans) to which just the virus, the antisera and the antibody were added and glycan-coated wells to which only the antisera and the antibody were added.

Ferret replication. We evaluated the replication kinetics of the α 2,3 H1N1pdm virus in the respiratory tract of 6–8-month-old male ferrets as previously

described¹¹. Briefly, all ferrets were screened before infection by HAI assay to ensure that they were naive to seasonal influenza A and B viruses. Animals were infected intranasally with 10^6 TCID₅₀ of α 2,3 H1N1pdm virus in 500 μl . Tissues were harvested to assess viral titres. Tissues were weighed and homogenized in Leibovitz's L-15 (L-15, Invitrogen) at 5% (nasal turbinates and trachea) or 10% (lung) weight per volume (W/V). The soft palate was homogenized in 1 ml of L-15. Clarified supernatant was aliquoted and titred on MDCK cells. The 50% tissue culture infectious dose (TCID₅₀) per gram of tissue was calculated by the Reed and Muench method³¹.

Influenza A virus full genome sequencing. The influenza A genomic RNA segments were simultaneously amplified from 3 μl of purified RNA (from homogenized ferret tissue) using a multi-segment RT-PCR strategy (M-RT-PCR)³². In a separate reaction, each HA segment was amplified using HA-specific primers (swH1ps-1A-F: 5'-AGCAAAAGCAGGGGAAAAACAAAGCAAC-3'; swH1ps-1777A-R: 5'-AGT AGAAACAAGGTGTTTCTCATGC-3'). Analysis of influenza viral RNA from ferret trachea and region of nasal turbinates enriched for respiratory epithelium, between the canine and second premolar teeth, was collected from tissue stored in RNAlater (Ambion) and total RNA was extracted using the RNeasy kit (Qiagen). For these samples, nested HA-specific small amplicons were generated using HA-specific PCR primers (outer primer pair H1-399F: 5'-AGCTCAGTGCATCA TTTGAAAG-3' and H1-961R: 5'-TGAAATGGGAGGCTGGTGTT-3'; and inner primer pair H1-468 F: 5'-AACAAAGGTGTAACGGCAGC-3' and H1-884R: 5'-AATGATAATACCAGATCCAGCAT-3'). Illumina libraries were prepared from M-RT-PCR products and from HA-specific RT-PCR products using the Nextera DNA Sample Preparation Kit (Illumina, Inc.) with half-reaction volumes.

After PCR amplification, 10 μl of each library derived from M-RT-PCR products was pooled into a 1.5 ml tube; separately, 10 μl of each library derived from HA-specific amplicons was pooled into a 1.5 ml tube. Each pool was cleaned two times with Ampure XP Reagent (Beckman Coulter) to remove all leftover primers and small DNA fragments. The first and second cleanings used $1.2 \times$ and $0.6 \times$ volumes of Ampure XP Reagent, respectively. The cleaned pool derived from M-RT-PCR products was sequenced on the Illumina HiSeq 2000 instrument (Illumina, Inc.) with 100-bp paired-end reads, while the cleaned pool derived from HA-specific amplicons was sequenced on the Illumina MiSeq v2 instrument with 300-bp paired-end reads. For additional sequencing coverage, and the HA-specific small amplicons, samples were re-sequenced using the Ion Torrent platform. M-RT-PCR products were sheared for 7 min, and Ion-Torrent-compatible bar-coded adapters were ligated to the sheared DNA using the Ion Xpress Plus Fragment Library Kit (Thermo Fisher Scientific, Waltham, MA, USA) to create 400-bp libraries. Libraries were pooled in equal volumes and cleaned with the Ampure XP Reagent. Quantitative PCR was performed on the pooled, barcoded libraries to assess the quality of the pool and to determine the template dilution factor for emulsion PCR. The pool was diluted appropriately and amplified on ion sphere particles (ISPs) during emulsion PCR on the Ion One Touch 2 instrument (Thermo Fisher Scientific). The emulsion was broken, and the pool was cleaned and enriched for template-positive ISPs on the Ion One Touch ES instrument (Thermo Fisher Scientific). Sequencing was performed on the Ion Torrent PGM using a 318v2 chip (Thermo Fisher Scientific).

Deep sequencing analysis. Deep sequencing preparation, collection and analysis were conducted by investigators who were blinded to the experimental groups. For virus sequence assembly, all sequence reads were sorted by barcode, trimmed, and *de novo* assembled using CLC Bio's *clc_novo_assemble* program (Qiagen, Hilden, Germany). The resulting contigs were searched against custom full-length influenza segment nucleotide databases to find the closest reference sequence for each segment. All sequence reads were then mapped to the selected reference influenza A virus segments using CLC Bio's *clc_ref_assemble_long* program.

Minor allele variants were identified using FindStatisticallySignificantVariants (FSSV) software (<http://sourceforge.net/projects/elvira/>). The FSSV software applies statistical tests to minimize false-positive SNP calls generated by Illumina sequence-specific errors (SSEs) described in ref. 33. SSEs usually result in false SNP calls if sequences are read in one sequencing direction. The FSSV analysis tool requires observing the same SNP at a statistically significant level in both sequencing directions. Once a minimum minor allele frequency threshold and significance level are established, the number of minor allele observations and major allele observations in each direction and the minimum minor allele frequency threshold are used to calculate *P*-values based on the binomial distribution cumulative probability. If the *P*-values calculated in both sequencing directions are less than the Bonferroni-corrected significance level, then the SNP calls are accepted. A significance level of 0.05 (Bonferroni-corrected for tests in each direction to 0.025) and a minimum minor allele frequency threshold of 3% were applied for this analysis. Differences in the consensus sequence compared to the reference sequence were identified using CLC Bio's *find_variations* software. The identified consensus and minor allele variations were analysed by assessing

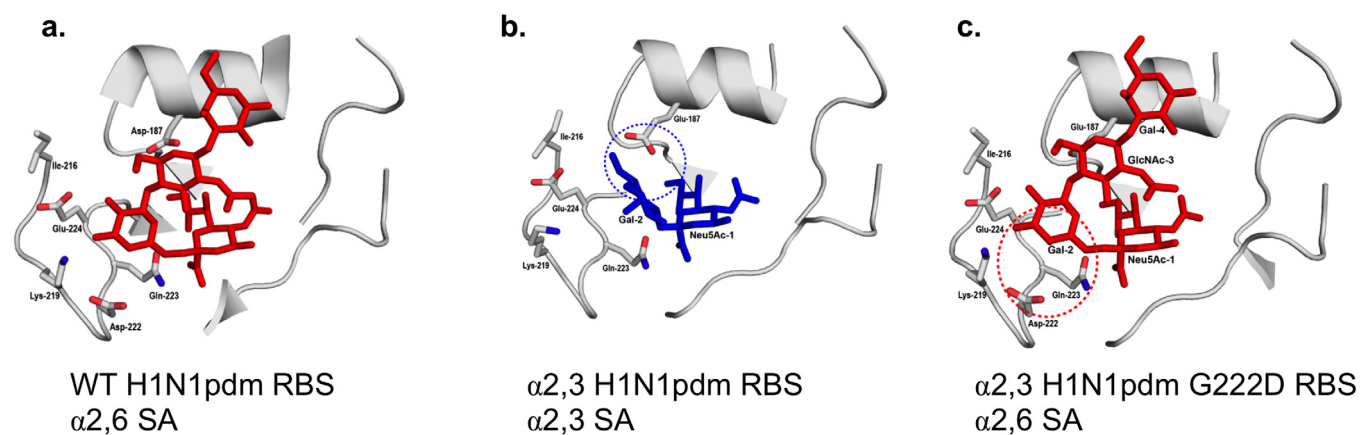
the functional impact on coding sequences or other regions based on overlap with identified features of the genome. For each sample, the reference sequence was annotated using VIGOR software³⁴, and then the variant data and genome annotation were combined using VariantClassifier software³⁵ to produce records describing the impacts of the identified variations.

Lectin and immunohistochemistry. Lectin histochemistry was performed as described previously for plant lectins³⁶ and purified HA protein³⁷. For plant lectin staining, the soft palate was subjected to microwave-based antigen retrieval using a citrate buffer and was then incubated with FITC-conjugated *Sambucus nigra* agglutinin (SNA) and biotinylated *Maackia amurensis* agglutinin (MAL II) lectins (Vector Laboratories), followed by a streptavidin-Alexa-Fluor594 conjugate (Invitrogen). For SC18 staining, the tissue sections were incubated with pre-complexed purified His-tagged SC18 HA protein, mouse anti-His antibody (Abcam), and goat anti-mouse IgG secondary antibody conjugated to Alexa-Fluor 488 (Molecular Probes) at a 4:2:1 ratio. Nuclei were counter stained with DAPI (Vector Laboratories) and sections were mounted with either ProLong Gold anti-fade reagent (Invitrogen) or Fluoromount-G (Southern Biotech). Images were captured either on an Olympus BX51 microscope with an Olympus DP80 camera or a Leica SP5 confocal microscope.

Ferret nasal turbinate biopsy samples were obtained from an uninfected 8 month old ferret as follows: the head was dissected sagittally to expose two halves of the ferret nasal turbinates; biopsy of turbinates between the canine and second

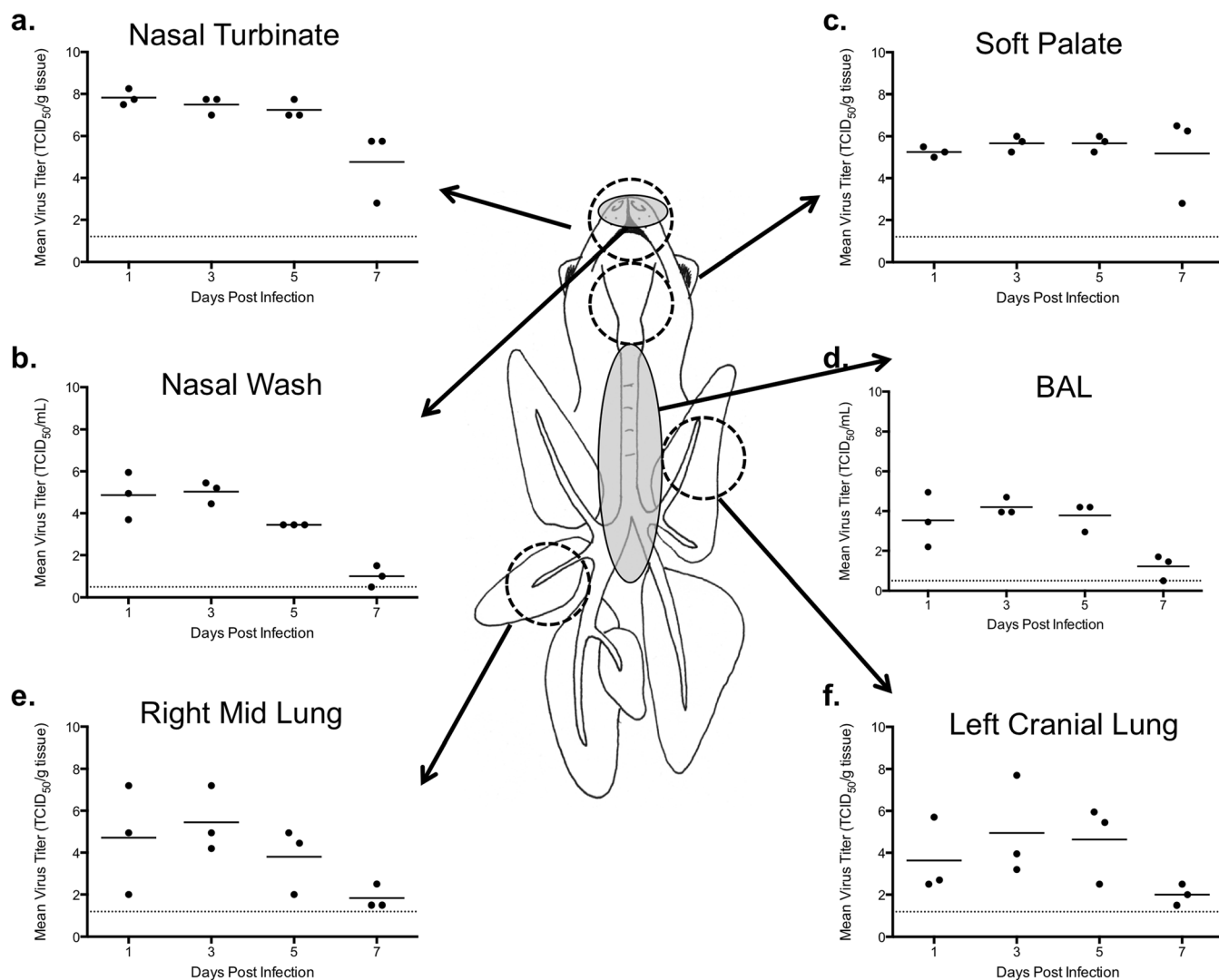
premolar represented respiratory epithelium and biopsy of turbinates at the molar tooth represented olfactory epithelium. A schematic depicting these two areas is shown in Extended Data Fig. 6h. Pig soft palate tissue sections were a gift from X. J. Meng and P. Pineyro. Pig soft palate tissues were collected from four 56-day-old mixed-breed commercial swine and fixed in 10% formalin. Soft palate tissues from four adult cadavers were obtained from the Maryland State Anatomy Board, Department of Health and Mental Hygiene in Baltimore, Maryland.

31. Reed, L. J. & Muench, H. A simple method of estimating fifty percent endpoints. *Am. J. Hyg.* **27**, 493–497 (1938).
32. Zhou, B. *et al.* Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza A viruses. *J. Virol.* **83**, 10309–10313 (2009).
33. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90 (2011).
34. Wang, S., Sundaram, J. P. & Stockwell, T. B. VIGOR extended to annotate genomes for additional 12 different viruses. *Nucleic Acids Res.* **40**, W186–W192 (2012).
35. Li, K. & Stockwell, T. B. VariantClassifier: A hierarchical variant classifier for annotated genomes. *BMC Res. Notes* **3**, 191 (2010).
36. Matsuoka, Y. *et al.* African green monkeys recapitulate the clinical experience with replication of live attenuated pandemic influenza virus vaccine candidates. *J. Virol.* **88**, 8139–8152 (2014).
37. Jayaraman, A. *et al.* Decoding the distribution of glycan receptors for human-adapted influenza A viruses in ferret respiratory tract. *PLoS ONE* **7**, e27517 (2012).



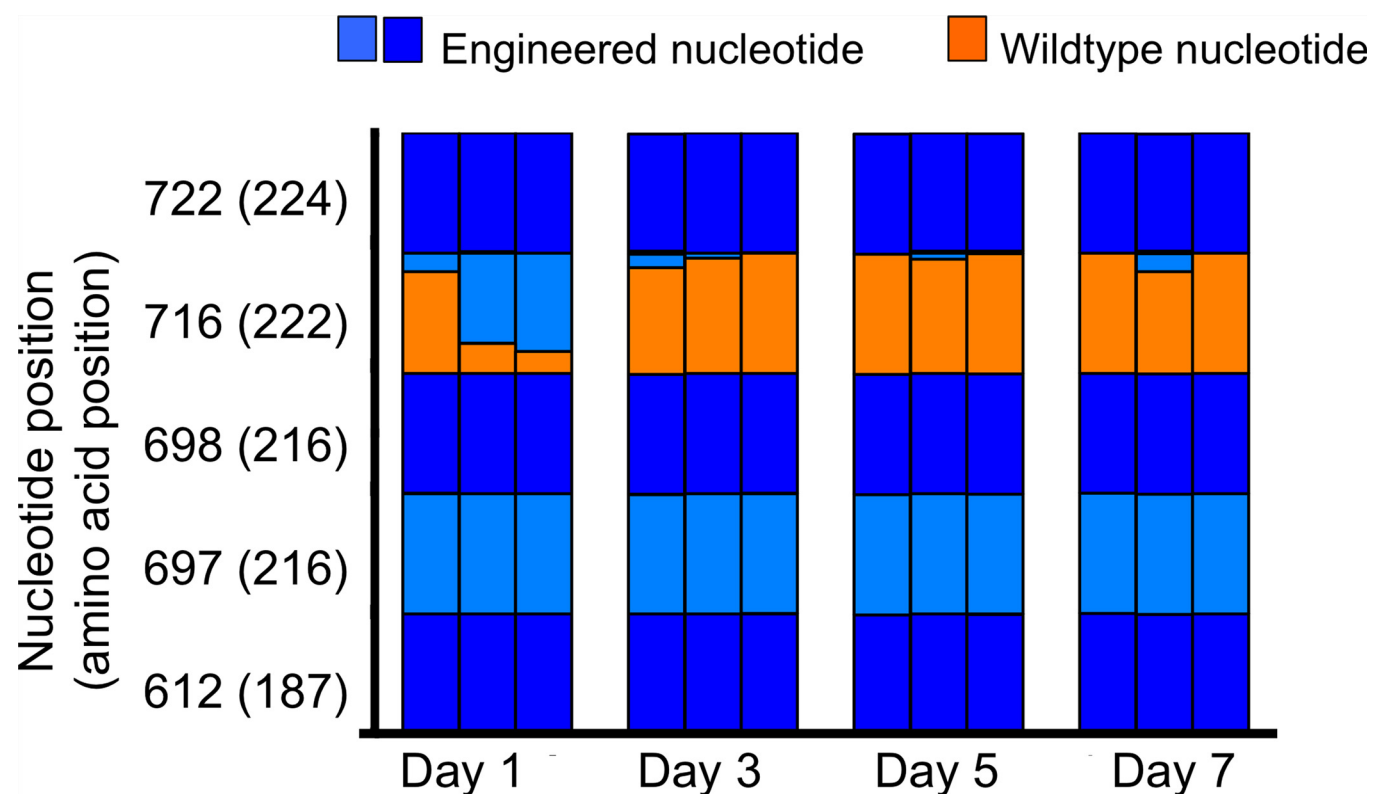
Extended Data Figure 1 | Amino acids in the receptor binding site of the 2009 H1N1pdm HA that bind to α2,3-linked and α2,6-linked sialic acid glycan receptor. a–c, Ribbon diagrams of the 2009 H1N1pdm HA receptor

binding pocket interacting with an α2,6-linked sialic acid glycan in the pocket (a), an α2,3 H1N1pdm HA with α2,3-linked sialic acid glycan (b), an α2,3 G222D revertant H1N1pdm HA with an α2,6-linked sialic acid glycan (c).



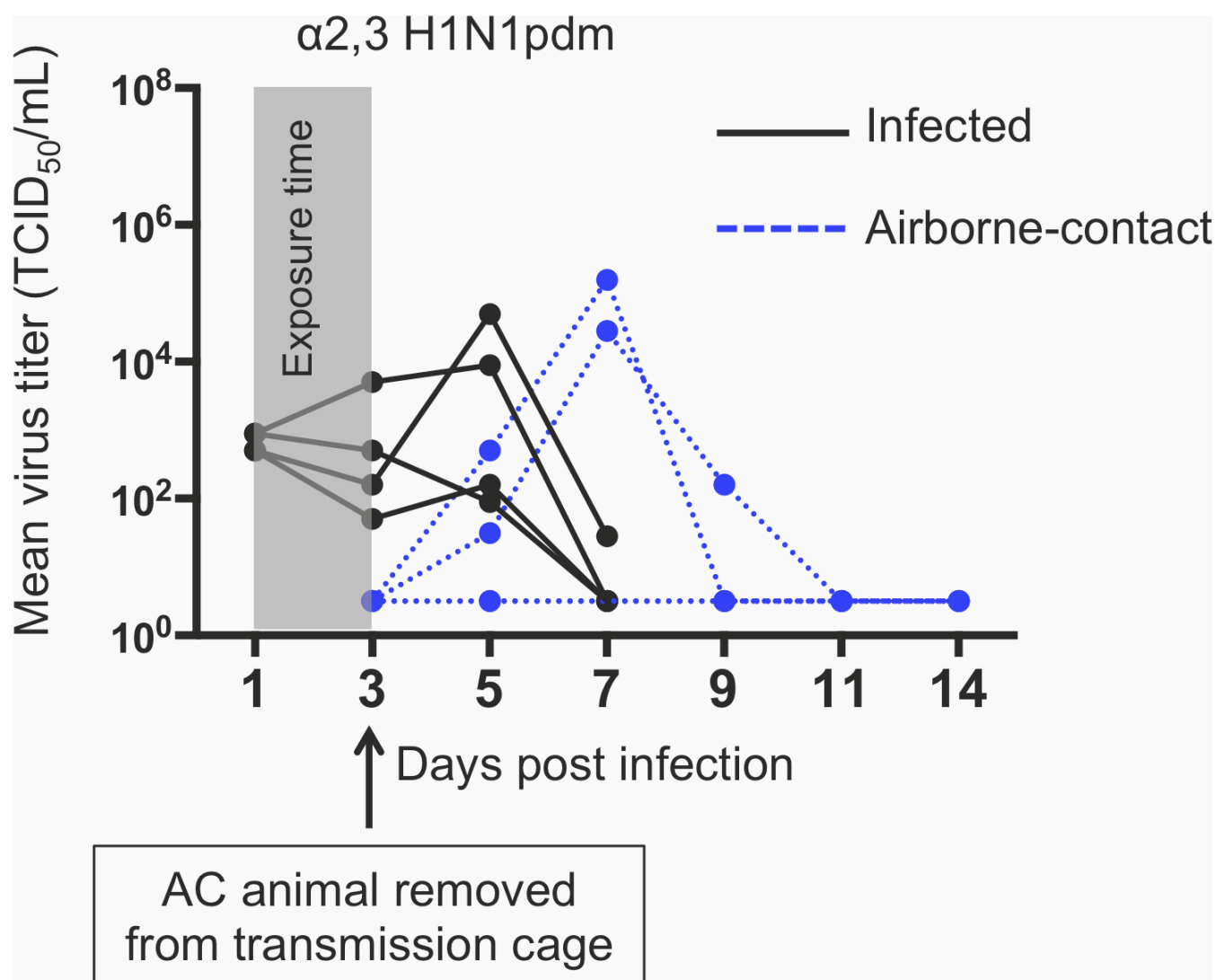
Extended Data Figure 2 | Replication of $\alpha 2,3$ H1N1pdm virus in ferret respiratory tract. a–f, We confirmed that the $\alpha 2,3$ H1N1pdm virus replicated to high titres on days 1, 3, 5 and 7 in different parts of the ferret respiratory tract.

Each tissue homogenate is highlighted with a dashed circle; the grey circles represent washes. Each point represents a single animal. The horizontal black line indicates the mean viral titre on a given day.



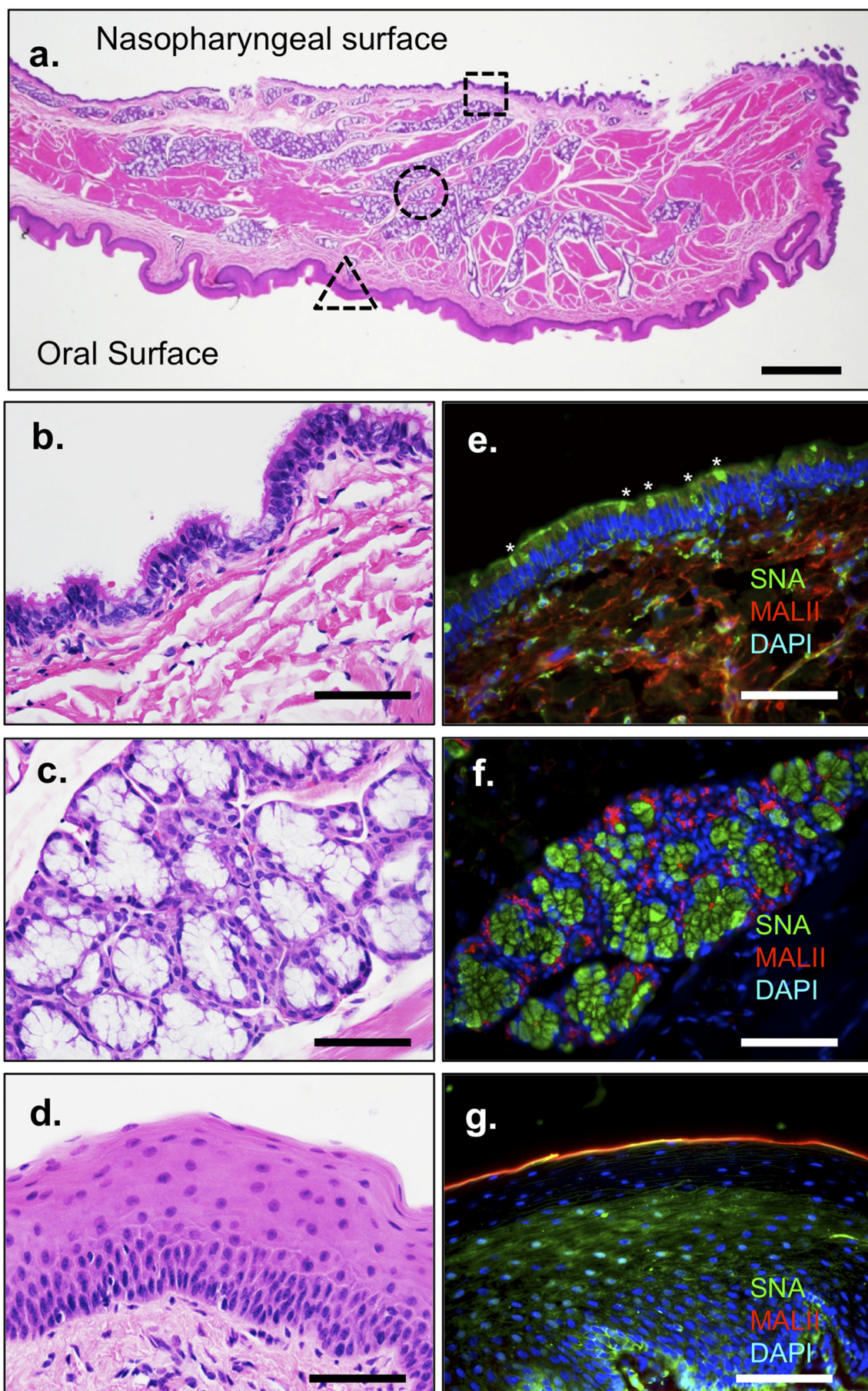
Extended Data Figure 3 | Stability of engineered mutations in viruses replicating in the soft palate. Deep sequencing of the HA gene segment from virus populations in the soft palate from 1, 3, 5 and 7 dpi reveals a rapid change

at position 222, but no change in the other engineered sites. The engineered sites are highlighted in blue, while the wild-type nucleotide is in orange. Each bar represents a single animal.



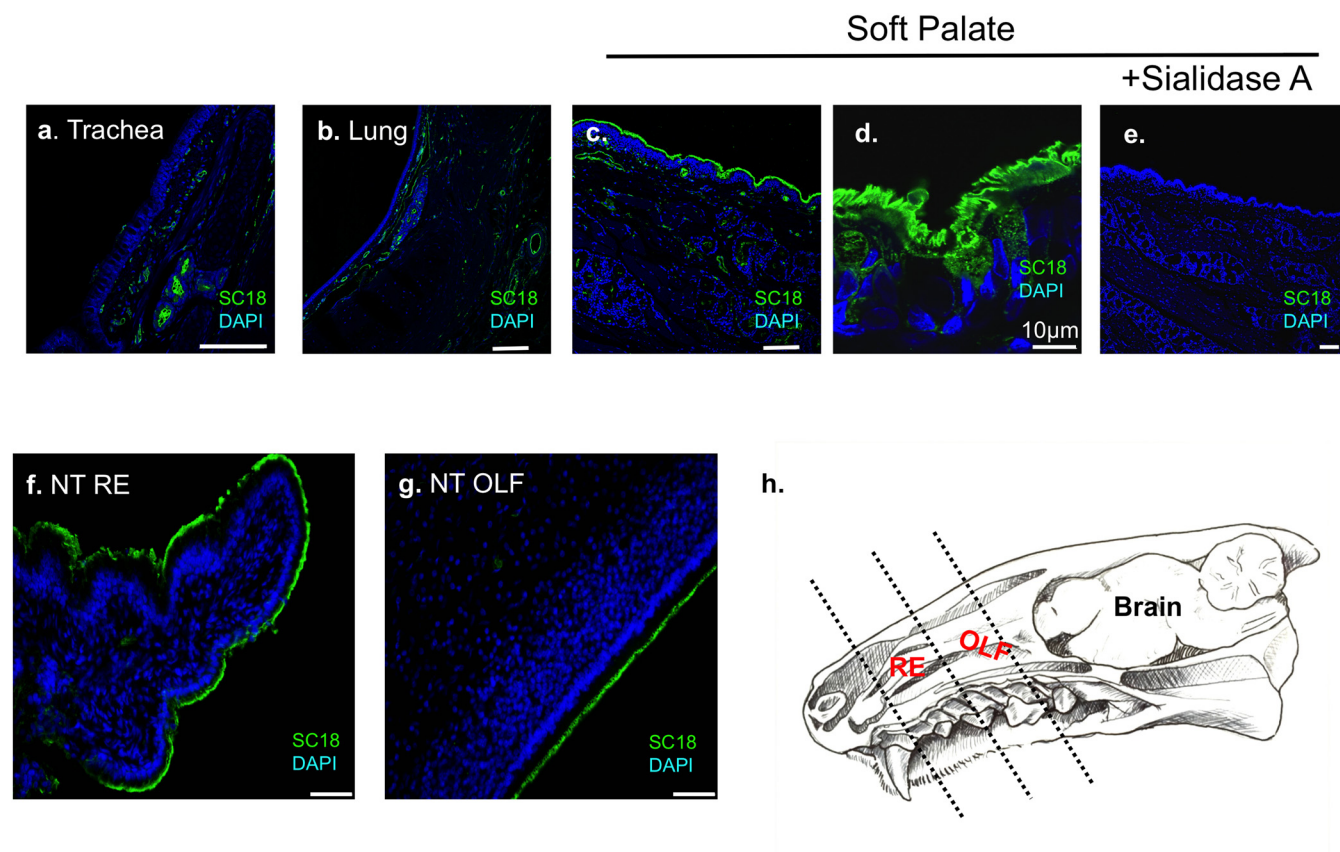
Extended Data Figure 4 | Airborne transmission of $\alpha 2,3$ H1N1pdm virus after 48 h exposure time. Transmission studies were performed with 4 pairs of animals (8 animals total) in double secure cages with perforated dividers. One ferret in each pair was infected with 10^6 TCID₅₀ of the indicated virus; a naive ferret (referred to as airborne-contact) was introduced into the adjacent compartment 24 h later. The airborne-contact animal was removed from the

transmission cage on day 3 post-infection as indicated by the black arrow. Nasal secretions were collected every other day for 14 days. Viral titres from the nasal secretions are graphed for each infected or airborne-contact animal. The grey shading indicates the exposure time between the infected and airborne-contact animals.



Extended Data Figure 5 | Influenza receptor distribution on ferret soft palate. Haematoxylin and eosin (H&E) staining of the soft palate from an uninfected ferret highlights the nasopharyngeal and oral surfaces. Scale bar, 1.25 mm. **a**, Areas highlighted in parts **b–g** are marked with dashed polygons: square, nasopharyngeal surface (**b** and **e**); circle, submucosal gland (**c** and

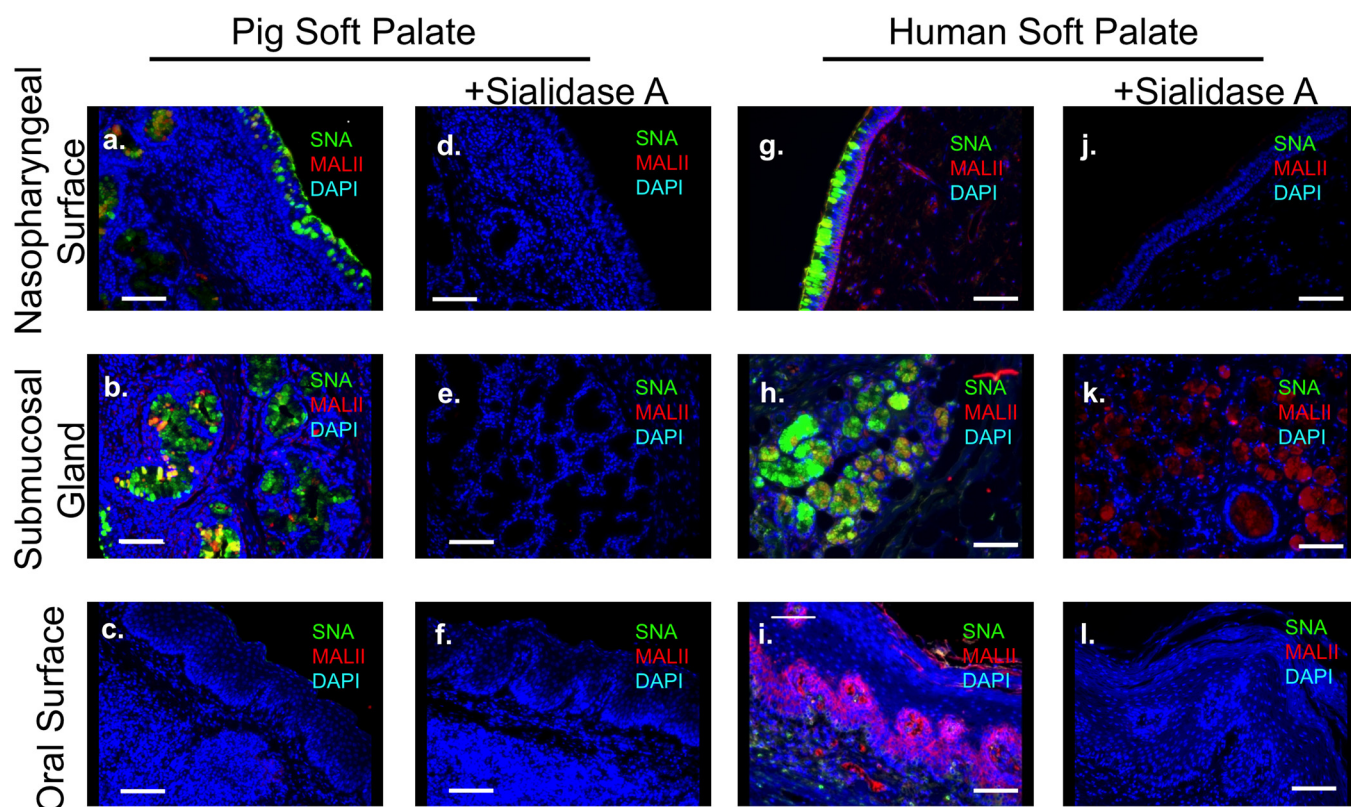
f); and triangle, oral surface (**d** and **g**). **b–d**, H&E staining of these regions, reproduced from Fig. 4a–c, are shown. **e–g**, Staining with plant lectins specific for α 2,6 sialic acid (SNA) and α 2,3 sialic acid (MAL II) are shown. Scale bars are 100 μ m in images **b–g**.



Extended Data Figure 6 | SC18 staining of ferret respiratory tissues.

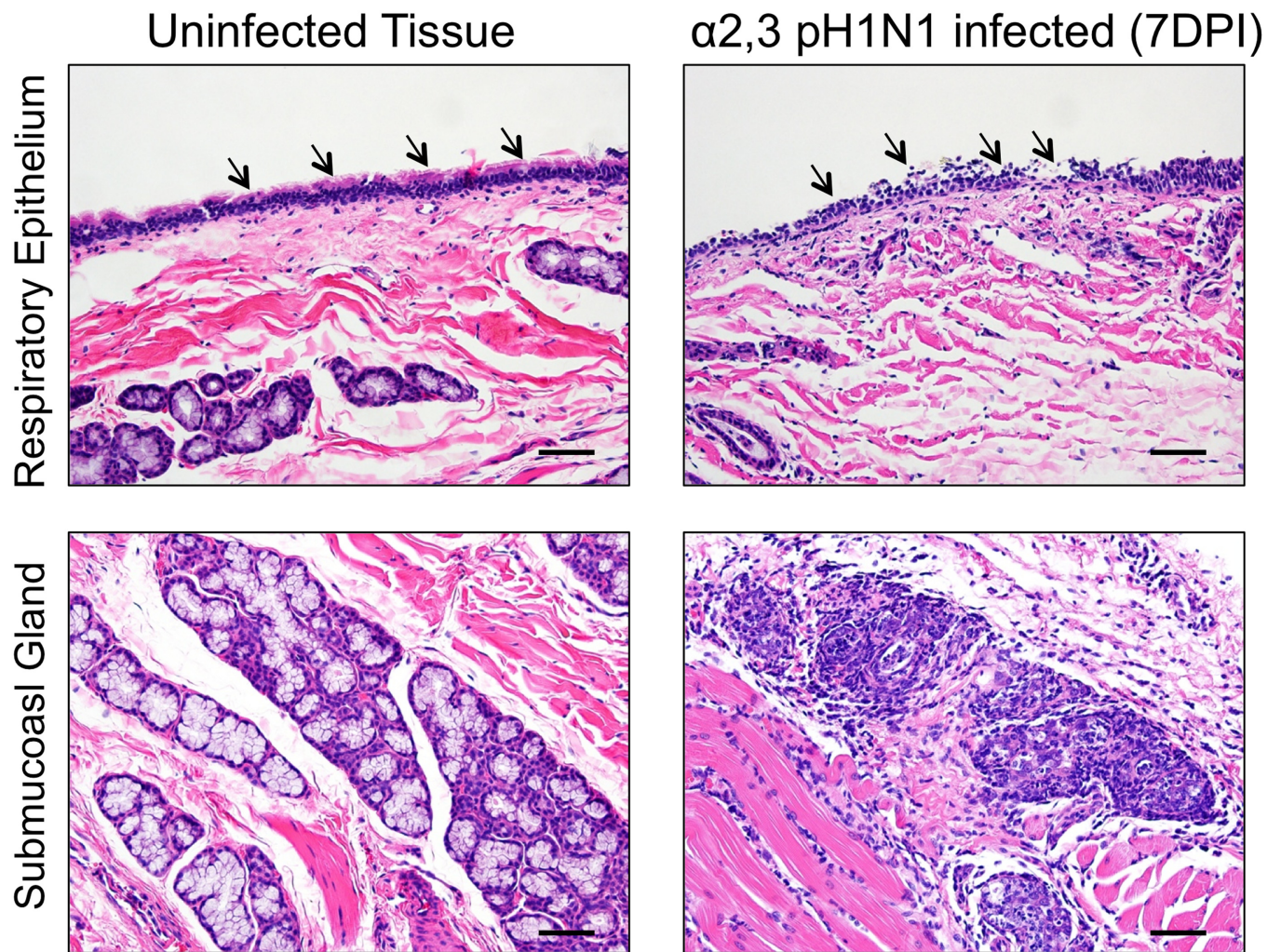
a–d, f, g. Sections of ferret trachea (a), lung (b), soft palate (c, d), biopsy of nasal turbinates (NT) tissue with respiratory epithelium (RE) (f) and olfactory epithelium (OLF) (g) were stained with purified SC18 HA protein to identify areas expressing long-chain $\alpha 2,6$ -linked sialic acids. **h.** Illustration of ferret head (sectioned along the midline) highlighting the anatomical locations of

respiratory epithelium and OLF tissues. Goblet cells on the respiratory epithelium of the soft palate (nasopharyngeal surface) also stained positive for SC18 (d, e). Absence of SC18 staining after sialidase A treatment indicates the high specificity of SC18 for the respiratory epithelium of the soft palate. All scale bars are 100 μ m unless indicated.



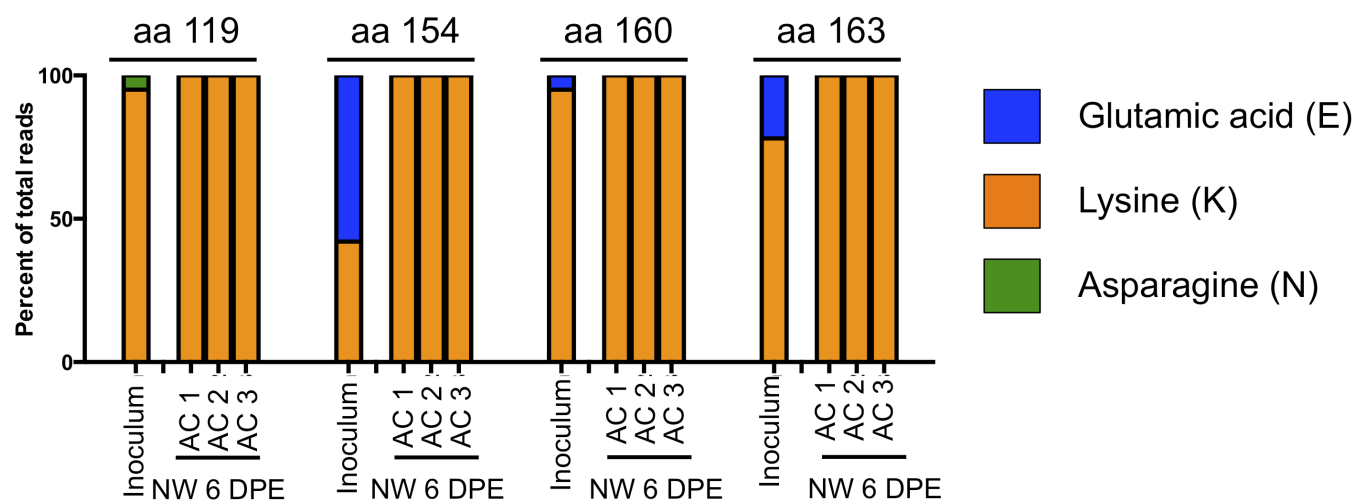
Extended Data Figure 7 | Influenza receptor distribution on pig and human soft palate. a–c, g–i, Pig (a–c) and human (g–i) soft palate tissues were stained with plant lectins SNA and MALII which are commonly used as markers for α 2,6 and α 2,3 sialic acids, respectively. d–f, j–l, Sialidase-A-treated control was run for each sample to ensure specificity of plant lectins and are displayed. Expression of α 2,6 sialic acids (SNA staining) is found on the ciliated respiratory epithelium and goblet cells of the nasopharyngeal surface and in the

submucosal glands of both the pig and human soft palate. Expression of α 2,3-linked sialic acids is low in the pig soft palate and found primarily in goblet cells and submucosal glands. In the human soft palate, MALII (α 2,3-linked sialic acids) staining sensitive to sialidase A treatment is found in the goblet cells and respiratory epithelium of the nasopharyngeal surface and in the basal cells of the oral surface. MALII staining in the submucosal glands was not sensitive to sialidase A treatment. Scale bars, 100 μ m.

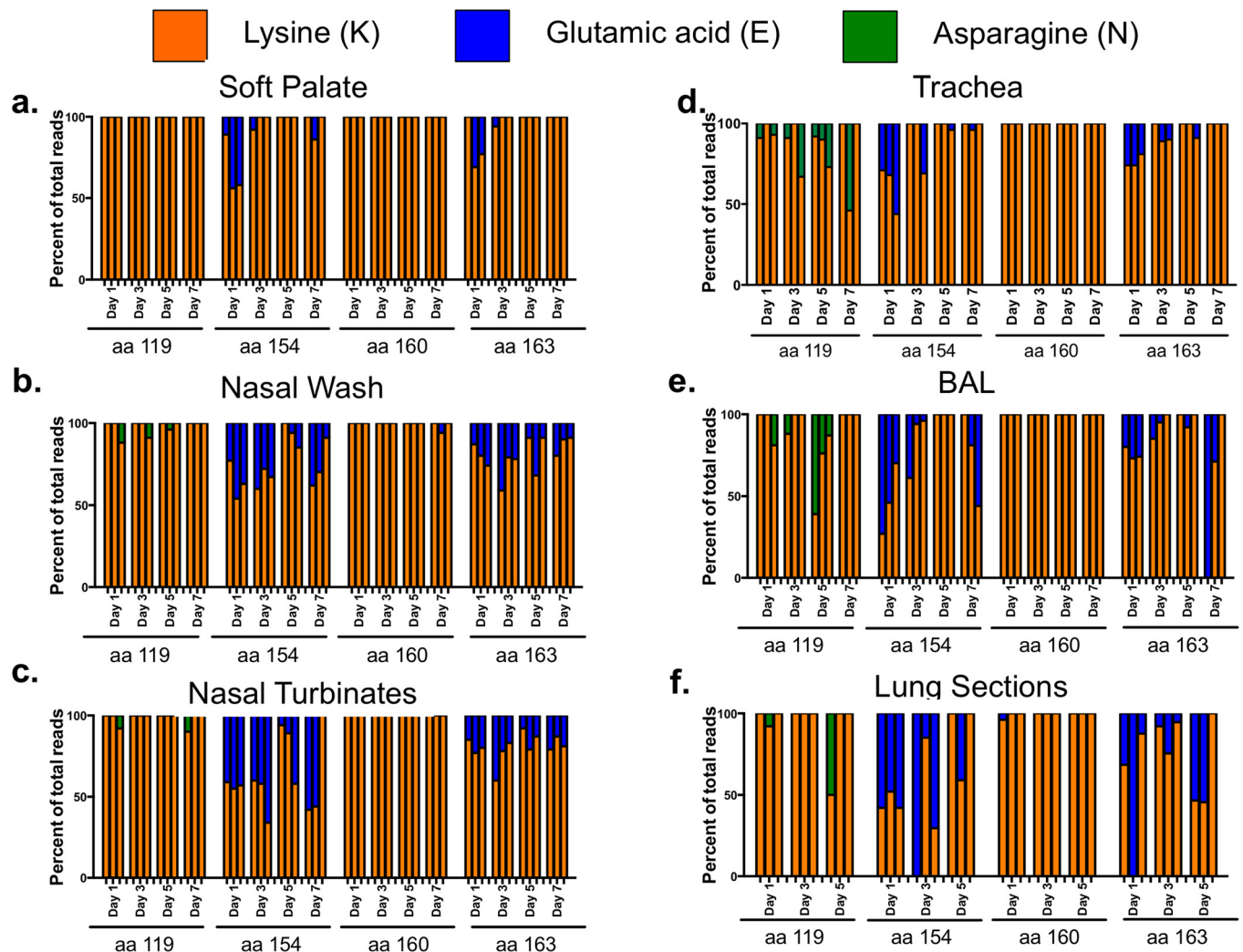


Extended Data Figure 8 | Pathology of the soft palate during infection with α2,3 H1N1pdm virus. The soft palate was removed from three ferrets infected with α2,3 H1N1pdm virus on 7 dpi. The tissue sections were stained with

haematoxylin and eosin. Black arrows indicate the ciliated respiratory epithelium of the soft palate tissue (nasopharyngeal surface). Scale bars, 100 μm in all images.



Extended Data Figure 9 | Quasi-species in putative lysine fence. Deep sequencing analysis of the α 2,3 H1N1pdm inoculum revealed a mixed population at four lysine residues surrounding the receptor binding site of the HA protein. The lysine fence was restored in viruses from the nasal wash of airborne-contact animals from 6 days post-exposure (DPE). Each bar represents a single animal, and each amino acid (aa) that contained a quasi-species is indicated.



Extended Data Figure 10 | Quasi-species of lysine fence in various ferret respiratory tissue sections. Deep sequencing of viruses from respiratory tissues of ferrets infected with $\alpha 2,3$ H1N1pdm virus. **a–f**, Virus populations from the soft palate (**a**), nasal wash (**b**), nasal turbinates (**c**), trachea

(**d**), bronchoalveolar lavage (BAL) (**e**), or lung sections (**f**) were analysed and the proportion of lysine, glutamic acid, or asparagine are presented. Each bar represents a single animal. The lung section is an average of the right middle lung lobe and a portion of the left caudal lung tissue.

Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal

Melih Acar², Kiranmai S. Kocherlakota^{1,2*}, Malea M. Murphy^{2*}, James G. Peyer^{2*}, Hideyuki Oguro², Christopher N. Inra², Christabel Jaiyeola², Zhiyu Zhao², Katherine Luby-Phelps³ & Sean J. Morrison^{1,2}

Haematopoietic stem cells (HSCs) reside in a perivascular niche but the specific location of this niche remains controversial¹. HSCs are rare and few can be found in thin tissue sections^{2,3} or upon live imaging⁴, making it difficult to comprehensively localize dividing and non-dividing HSCs. Here, using a green fluorescent protein (GFP) knock-in for the gene *Cttnl1* in mice (hereafter denoted as α -catulin^{GFP}), we discover that α -catulin^{GFP} is expressed by only 0.02% of bone marrow haematopoietic cells, including almost all HSCs. We find that approximately 30% of α -catulin^{GFP}+c-kit⁺ cells give long-term multilineage reconstitution of irradiated mice, indicating that α -catulin^{GFP}+c-kit⁺ cells are comparable in HSC purity to cells obtained using the best markers currently available. We optically cleared the bone marrow to perform deep confocal imaging, allowing us to image thousands of α -catulin^{GFP}+c-kit⁺ cells and to digitally reconstruct large segments of bone marrow. The distribution of α -catulin^{GFP}+c-kit⁺ cells indicated that HSCs were more common in central marrow than near bone surfaces, and in the diaphysis relative to the metaphysis. Nearly all HSCs contacted leptin receptor positive (Lepr⁺) and Cxcl12^{high} niche cells, and approximately 85% of HSCs were within 10 μ m of a sinusoidal blood vessel. Most HSCs, both dividing (Ki-67⁺) and non-dividing (Ki-67⁻), were distant from arterioles, transition zone vessels, and bone surfaces. Dividing and non-dividing HSCs thus reside mainly in perisinusoidal niches with Lepr⁺Cxcl12^{high} cells throughout the bone marrow.

Adult HSCs reside in a perivascular niche in the bone marrow in which Lepr⁺ perivascular stromal cells and endothelial cells secrete factors that promote their maintenance^{5–9}. Nearly all of the cells that express high levels of *Scf* (also known as *Kitl*) or *Cxcl12* in the bone marrow are Lepr⁺ (ref. 10). Conditional deletion of *Scf* from Lepr⁺ cells and endothelial cells eliminates all quiescent and serially transplantable HSCs from adult bone marrow¹¹. The perivascular niche cells we identified based on Lepr expression have also been identified by others based on their expression of high levels of *Cxcl12* (refs 5, 12, and 13) and low levels of the *Nes*–GFP transgene^{14,15}, PDGFR^{10,16}, and *Prx1*–Cre⁸.

Established elements of the HSC niche localize primarily around sinusoids in bone marrow including HSCs^{2,3,17}, Lepr⁺ stromal cells⁶, angiopoietin-1-expressing stromal cells^{18,19}, *Scf*-expressing stromal cells⁶, Cxcl12^{high} stromal cells^{5,12,20}, and mesenchymal stem/stromal cells^{10,18,20}. Moreover, sinusoidal endothelial cells are functionally important for haematopoiesis after myeloablation²¹. HSCs have also been suggested to reside in a hypoxic niche²² and the most hypoxic region of the bone marrow is around sinusoids²³. Nonetheless, HSC niches have also been suggested to localize near bone surfaces or around arterioles²⁴.

It has been suggested that dividing HSCs reside in a niche that is spatially distinct from quiescent HSCs²⁴. However, dividing HSCs are rarer than non-dividing HSCs, making it difficult to find substantial

numbers of those cells within tissue sections. HSC imaging throughout the bone marrow is limited by the inability of even multiphoton microscopy to penetrate more than 150 μ m into bone marrow⁴. Optical clearing techniques have enabled deep imaging of various tissues^{25,26}, including haematopoietic progenitors in embryos²⁷, but have not been used to image rare stem cells or to digitally reconstruct bone marrow.

Gene expression profiling showed that α -catulin is highly restricted in its expression to HSCs³. α -catulin encodes a protein with homology to α -catenin that has been suggested to function as a cytoskeletal linker²⁸. By quantitative reverse transcription PCR (qRT–PCR), we found that α -catulin was expressed at 19 ± 9.3 (mean \pm s.d.) fold higher levels in CD150⁺CD48⁻Lineage⁻Sca-1⁺c-kit⁺ (CD150⁺CD48⁻LSK) HSCs as compared to unfractionated bone marrow cells.

To assess α -catulin expression in detail, we knocked GFP into the first exon of α -catulin in frame with the start codon (Extended Data Fig. 1a). Although this was predicted to be a loss-of-function allele, both α -catulin^{GFP/+} and α -catulin^{GFP/GFP} mice were born and survived into adulthood with expected Mendelian frequencies (Extended Data Fig. 1e). Young adult α -catulin^{GFP/GFP} mice were normal in size and body mass (Extended Data Fig. 1d), as well as bone density and bone volume (Extended Data Fig. 1f) relative to littermate controls. α -catulin^{GFP/+} and α -catulin^{GFP/GFP} mice exhibited normal haematopoiesis as well as normal HSC frequency, HSC cell cycle kinetics, and normal HSC function upon primary and secondary transplantation into irradiated mice (Extended Data Fig. 2).

Only $0.021 \pm 0.006\%$ (mean \pm s.d.) of whole bone marrow (WBM) cells in α -catulin^{GFP/+} mice were α -catulin^{GFP}+ (Fig. 1a). Most of the α -catulin^{GFP}+ cells were also c-kit⁺ (Extended Data Fig. 3a), and $49 \pm 8.3\%$ of CD150⁺CD48⁻LSK HSCs were α -catulin^{GFP}+ (Extended Data Fig. 3c). We did not detect α -catulin^{GFP} expression among multipotent progenitors, common lymphoid progenitors, common myeloid progenitors, granulocyte macrophage progenitors, or megakaryocyte erythrocyte progenitors (Extended Data Fig. 3c, d). α -catulin^{GFP}+c-kit⁺ cells seemed to be highly purified HSCs as they represented only $0.007 \pm 0.003\%$ of WBM cells and were uniformly CD150⁺ and CD48⁻ (Fig. 1b and Extended Data Fig. 3b).

To test the function of α -catulin^{GFP}+ cells, we performed long-term competitive reconstitution assays in irradiated mice. We found that 1 in 37,000 WBM cells gave long-term multilineage reconstitution; and 1 in 6.7 α -catulin^{GFP}+ cells gave long-term multilineage reconstitution (Table 1). In contrast, only 1 in 2,847,000 α -catulin^{GFP}- bone marrow cells gave long-term multilineage reconstitution (a 77-fold depletion over WBM). The α -catulin^{GFP}+ fraction of CD150⁺CD48⁻LSK cells gave long-term multilineage reconstitution (1 in 3.1 cells) but the α -catulin^{GFP}- fraction had little HSC activity (1 in 110 cells; Table 1). Therefore, nearly all HSC activity in adult bone marrow is contained within the α -catulin^{GFP}+ fraction of cells.

¹Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. ²Children's Research Institute and the Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. ³Department of Cell Biology, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.

*These authors contributed equally to this work.

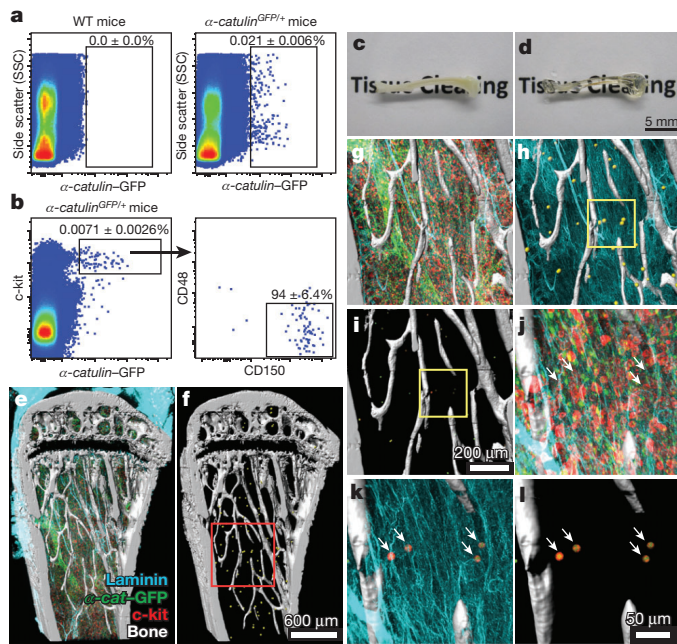


Figure 1 | Deep imaging of α -catulin-GFP⁺ HSCs in digitally reconstructed bone marrow. **a**, Only $0.021 \pm 0.006\%$ of α -catulin^{GFP/+} bone marrow cells were GFP⁺ ($n = 14$ mice in 11 independent experiments). **b**, Nearly all α -catulin-GFP⁺c-kit⁺ bone marrow cells were CD150⁺CD48⁻ ($n = 9$ mice in 3 independent experiments; Extended Data Fig. 3b shows ungated cells from this analysis and Extended Data Fig. 5 shows light scatter properties). **c, d**, A half tibia before (**c**) and after (**d**) clearing. **e–l**, Deep imaging of α -catulin-GFP⁺c-kit⁺ HSCs in the epiphysis and metaphysis of a half tibia (360 μ m thick) showing digital bone surfaces (second harmonic generation, white), as well as blood vessels (laminin, blue), haematopoietic progenitors (c-kit⁺, red), and α -catulin-GFP⁺ cells (green). Endothelial cells express α -catulin-GFP but were easily distinguished from α -catulin-GFP⁺c-kit⁺ HSCs based on c-kit expression and morphology. In 2D projected images of thick specimens, α -catulin-GFP⁺ cells and c-kit⁺ cells can appear more frequent than they actually are because all of the cells from the thick specimens are collapsed into a single optical plane. **f**, Same as **e**, but digitally masked to reveal only HSCs and bone. α -catulin-GFP⁺c-kit⁺ HSCs are represented by yellow spheres to make them visible at this magnification. **g–i**, Higher magnification views of the boxed region in **f**. **h**, All haematopoietic cells other than HSCs (yellow spheres) are digitally masked. **i**, Blood vessels and haematopoietic cells other than HSCs are digitally masked. **j–l**, Higher magnification views of the boxed area from **i** (arrows indicate α -catulin-GFP⁺c-kit⁺ cells). Images are representative of three independent experiments. Supplementary Video 1 shows a 3D digital reconstruction of bone and bone marrow. The positions of HSCs and other structures can appear to change in thick specimens when magnification is changed due to the rendering perspective for 3D display of volume data.

Furthermore, 1 in 3.5 α -catulin-GFP⁺c-kit⁺ cells gave long-term multilineage reconstitution of irradiated primary and secondary recipient mice (Table 1; Extended Data Fig. 4). If HSCs home with near perfect efficiency to engraft haematopoietic tissues, then approximately 30% of α -catulin-GFP⁺c-kit⁺ cells are HSCs. If most intravenously transplanted HSCs fail to home to haematopoietic tissues and therefore do not have an opportunity to exhibit reconstituting potential in transplantation assays, then most α -catulin-GFP⁺c-kit⁺ cells may be HSCs. In either case, α -catulin-GFP⁺c-kit⁺ cells are comparable in purity to cells isolated using the best HSC markers currently available.

α -catulin-GFP⁺c-kit⁺ cells are quiescent, comparable to CD150⁺CD48⁻LSK HSCs, with only $1.2 \pm 0.5\%$ of cells in S/G2/M phases of the cell cycle (Extended Data Fig. 4c).

To systematically identify all of the α -catulin-GFP⁺c-kit⁺ HSCs within a large segment of bone marrow, we implemented a clearing technique^{25,26} which permitted deep confocal imaging. After antibody

Table 1 | α -catulin-GFP⁺ cells were highly enriched for long-term multilineage reconstituting (LTMR) HSCs.

| Donor cell population | Donor cell dose | LTMR mice/transplanted mice | HSC frequency |
|--|--|----------------------------------|----------------|
| Unfractionated bone marrow | 25,000 50,000 100,000 300,000 | 19/27 15/25 21/23 14/14 | 1 in 37,000 |
| α -catulin-GFP ⁺ | 1 5 | 2/13 13/25 | 1 in 6.7 |
| α -catulin-GFP ⁻ | 300,000 | 2/20 | 1 in 2,847,000 |
| α -catulin-GFP ⁺ c-kit ⁺ | 1 5 30 | 9/34 17/23 14/14 | 1 in 3.5 |
| α -catulin-GFP ⁻ CD150 ⁺ CD48 ⁻ Lin ⁻ c-kit ⁺ Sca-1 ⁺ | 5 | 2/44 | 1 in 110 |
| α -catulin-GFP ⁺ CD150 ⁺ CD48 ⁻ Lin ⁻ c-kit ⁺ Sca-1 ⁺ | 5 | 24/30 | 1 in 3.1 |

Donor cells were competed against 300,000 recipient WBM cells in irradiated mice. The data reflect means from 2 to 4 independent experiments per cell population. LTMR mice indicates the number of mice that were long-term multilineage reconstituting as a fraction of the total number transplanted with each cell dose.

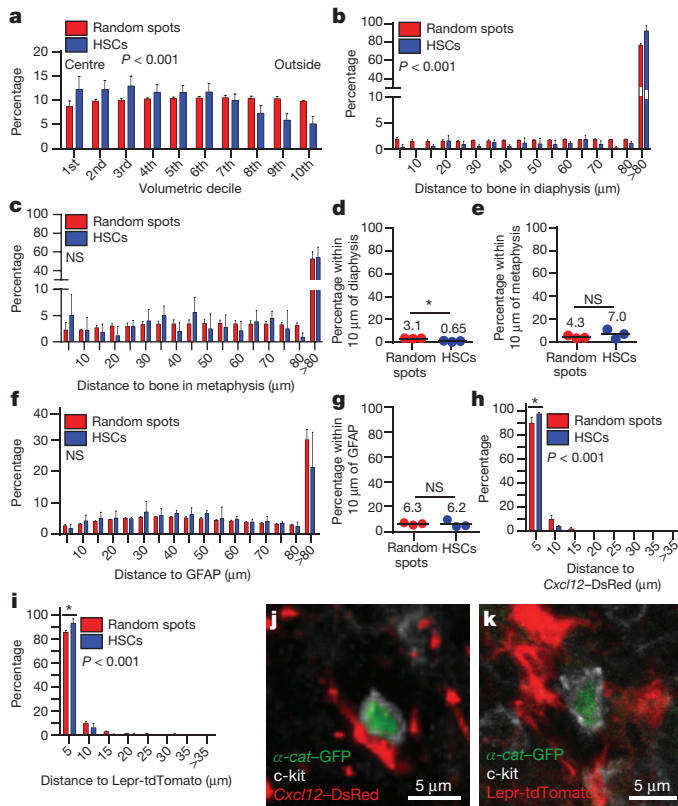
staining of half bones or bone marrow plugs from α -catulin^{GFP/+} mice, we cleared the specimens (see Fig. 1c versus d) and then used confocal microscopes to acquire tiled Z-stacked optical sections throughout the bone marrow to a depth of up to 600 μ m.

We identified all α -catulin-GFP⁺c-kit⁺ cells within large segments of bone marrow (Fig. 1e–l). Isotype controls showed low levels of background fluorescence that could readily be distinguished from positive signals (Extended Data Fig. 8a, b). We also prepared videos that animate three-dimensional (3D) images of the bone marrow to show HSCs, niche cells, vasculature, and bone surfaces (Supplementary Videos 1–3). When thick specimens are collapsed into a single 2D image, α -catulin-GFP⁺ cells and c-kit⁺ cells can appear more frequent than they actually are because all of the cells from the thick specimens are collapsed into a single 2D image (for example, see Extended Data Fig. 7i versus 7j). There are also cases in which an α -catulin-GFP⁺ cell and a c-kit⁺ cell are present in different optical planes but appear to be a single α -catulin-GFP⁺c-kit⁺ cell when collapsed into a single 2D image. For this reason, α -catulin-GFP⁺c-kit⁺ cells were individually examined at high magnification in 3D to confirm double labelling of single cells.

To analyse the location of α -catulin-GFP⁺c-kit⁺ HSCs relative to bone surfaces, we divided the bone marrow in the tibia diaphysis (shaft) into concentric cylindrical volumes that each encompassed 10% of the marrow volume (Extended Data Fig. 6a). Although HSCs were found throughout the marrow, they were significantly enriched towards the centre of the marrow and depleted near the bone surface (Fig. 2a and Extended Data Fig. 6g).

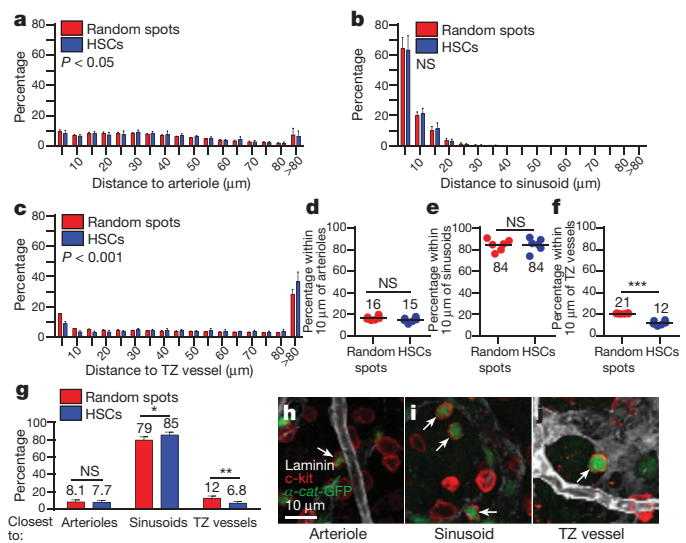
The frequencies of CD150⁺CD48⁻LSK HSCs and α -catulin-GFP⁺c-kit⁺ HSCs were both significantly lower in the epiphysis and metaphysis as compared to the diaphysis by flow cytometry in the femur and tibia (using WBM from crushed bones; Extended Data Fig. 7a, c, f, g). Consistent with this, 3D confocal imaging of bisected tibias showed the frequency of α -catulin-GFP⁺c-kit⁺ HSCs was significantly lower in the metaphysis as compared to the diaphysis (Extended Data Fig. 7h). Thus, both deep imaging and flow cytometric analysis indicated that HSCs are more enriched in the diaphysis than in the metaphysis.

In the diaphysis, α -catulin-GFP⁺c-kit⁺ HSCs were significantly less likely than randomly placed spots to localize close to a bone surface (Fig. 2b; random spots were distributed throughout areas occupied by haematopoietic cells but excluded from areas occupied by bone or blood vessel lumens). Only $0.65 \pm 0.58\%$ of HSCs were within



10 μ m of a bone surface in the diaphysis (Fig. 2d), and 86 \pm 6.1% were more than 80 μ m away (Fig. 2b). In the metaphysis, the localization of HSCs relative to bone surfaces did not significantly differ from the distribution of random spots as 7.0 \pm 4.1% of HSCs were within 10 μ m of a bone surface (Fig. 2e), and 53 \pm 11% were more than 80 μ m away (Fig. 2c). These data are consistent with our observation that fewer than 10% of CD150⁺CD48⁺LSK HSCs were within 10 μ m of bone in femur sections^{2,3}.

Glial fibrillary acidic protein (GFAP)⁺ Schwann cells and nerve fibres clustered in the centre of the marrow in the diaphysis (Extended Data Fig. 8d, e). HSCs did not significantly differ from random spots in their distance from GFAP⁺ cells (Fig. 2f). Only



6.2 \pm 3.0% of α -catulin-GFP⁺ c-kit⁺ HSCs were within 10 μ m of GFAP⁺ cells in the bone marrow but 28 \pm 3.8% were within 30 μ m (Fig. 2f, g). Thus, HSCs and niche cells rarely have contact with Schwann cells or nerve fibres but a subset of HSCs may be regulated by diffusible factors secreted by Schwann cells²⁹.

HSCs were significantly more likely than random spots to be close to *Cxcl12*-DsRed^{high} stromal cells, and 97 \pm 1.2% of HSCs were within 5 μ m of *Cxcl12*-DsRed^{high} stromal cells (Fig. 2h). Although *Cxcl12*^{high} stromal cells represent only 0.3% of WBM cells¹⁰, they have long processes that extend throughout the marrow (Supplementary Video 2). Consequently, 89 \pm 4.9% of random spots were also within 5 μ m of a *Cxcl12*-DsRed^{high} stromal cell (Fig. 2h), and 94 \pm 2.5% of HSCs appeared to have cell–cell contact with a *Cxcl12*-DsRed^{high} stromal cell (Fig. 2j).

We also found that 93 \pm 3.7% of α -catulin-GFP⁺ c-kit⁺ HSCs were within 5 μ m of a *Lepr*⁺ cell (Fig. 2i). *Lepr*⁺ cells were visualized using *Lepr-cre;tdTomato* mice in these experiments, but 99% of tdTomato⁺ bone marrow cells in 8–12 week old *Lepr-cre;tdTomato* mice also stain with an antibody against *Lepr*¹⁰. HSCs were significantly more likely than random spots to be close to (Fig. 2i) and almost always contacted (Fig. 2k) a *Lepr*⁺ cell.

We next imaged the localization of HSCs relative to three kinds of blood vessels in the bone marrow: arterioles, sinusoids, and transition zone capillaries³⁰. We distinguished blood vessels based on anatomical position, size, morphology, and continuity of the basal lamina, visualized using anti-laminin antibody staining (Extended Data Fig. 9a–c). α -catulin-GFP⁺ c-kit⁺ HSCs significantly differed from random spots in their distance to arterioles; they were slightly less likely than random spots to be within 25 μ m of an arteriole but slightly more likely than random spots to be 30–50 μ m away (Fig. 3a). Only 15 \pm 2.3% of HSCs were within 10 μ m of an arteriole (Fig. 3d). In contrast, 84 \pm 6.2% of HSCs were within 10 μ m of a sinusoid (Fig. 3e). HSCs did not significantly differ from random spots in their localization

10 μ m of a bone surface in the diaphysis (Fig. 2d), and 86 \pm 6.1% were more than 80 μ m away (Fig. 2b). In the metaphysis, the localization of HSCs relative to bone surfaces did not significantly differ from the distribution of random spots as 7.0 \pm 4.1% of HSCs were within 10 μ m of a bone surface (Fig. 2e), and 53 \pm 11% were more than 80 μ m away (Fig. 2c). These data are consistent with our observation that fewer than 10% of CD150⁺CD48⁺LSK HSCs were within 10 μ m of bone in femur sections^{2,3}.

Glial fibrillary acidic protein (GFAP)⁺ Schwann cells and nerve fibres clustered in the centre of the marrow in the diaphysis (Extended Data Fig. 8d, e). HSCs did not significantly differ from random spots in their distance from GFAP⁺ cells (Fig. 2f). Only

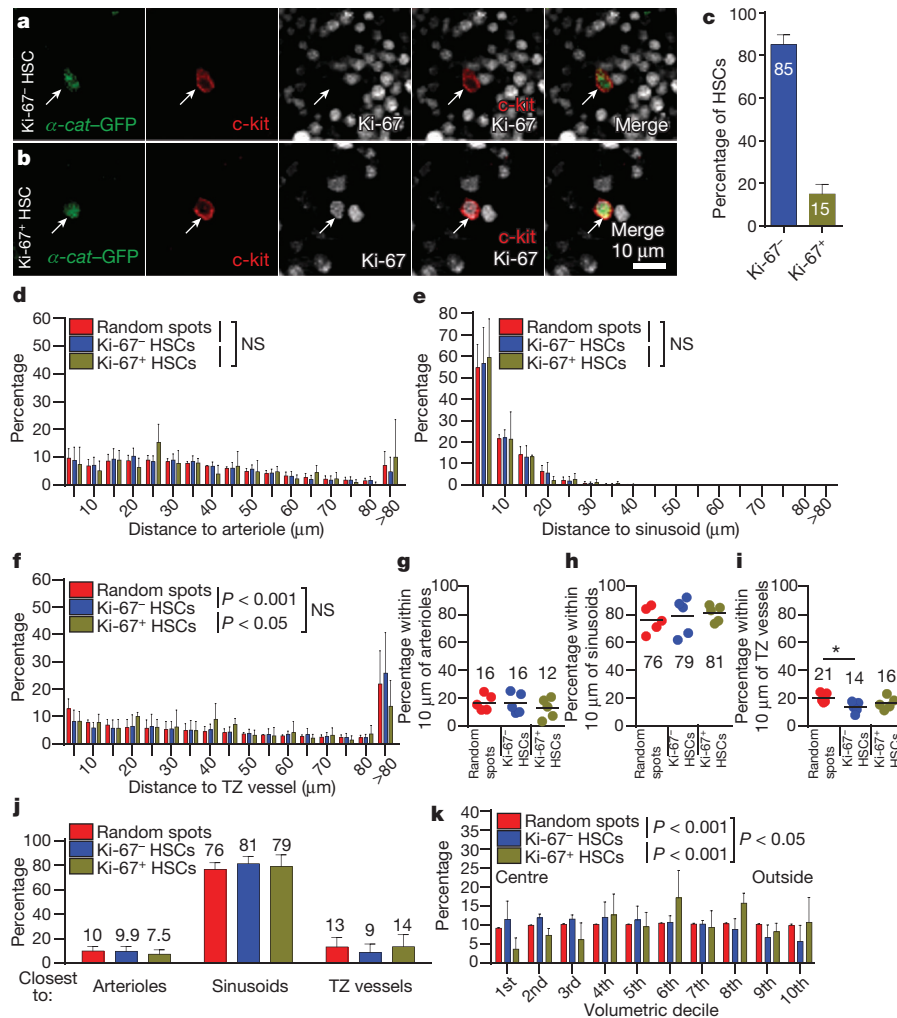


Figure 4 | Dividing and non-dividing HSCs are most closely associated with sinusoids. **a, b**, Representative images of a Ki-67⁻ α -catulin-GFP⁺*c-kit*⁺ non-dividing HSCs (**a**) and a Ki-67⁺ α -catulin-GFP⁺*c-kit*⁺ dividing HSCs (**b**) (arrows). **c**, 15 \pm 2.0% of HSCs were Ki-67⁺. All data reflect mean \pm s.d. from bone marrow plugs (410–440 μ m thick) from the diaphysis of 5 tibias. A total of 2,132 HSCs were analysed in 5 independent experiments. **d–f**, Distance to the nearest arteriole (**d**), sinusoid (**e**), or transition zone (TZ) vessel (**f**). **g–i**, The percentages of all Ki-67⁻ non-dividing HSCs, Ki-67⁺ dividing HSCs, or random spots within 10 μ m of an arteriole (**g**), a sinusoid (**h**), or a transition zone vessel (**i**). **j**, Most Ki-67⁻ non-dividing HSCs and Ki-67⁺ dividing HSCs were most closely

associated with sinusoids. **k**, The distributions of Ki-67⁻ non-dividing HSCs, Ki-67⁺ dividing HSCs, and random spots in concentric cylinders corresponding to equal volumetric deciles from central marrow (centre) to the marrow near the endosteal bone surface (outside). Non-dividing HSCs were significantly enriched in central marrow, whereas dividing HSCs were significantly enriched towards the endosteum (**d–k** reflect mean \pm s.d. from bone marrow plugs from the diaphysis of 5 tibias. A total of 1,840 Ki-67⁻ HSCs and 292 Ki-67⁺ HSCs were analysed in 5 independent experiments). In **d–f** and **k**, the statistical significance was assessed by Kolmogorov–Smirnov analysis, and in **g–j**, by Student's *t*-tests (* $P < 0.05$); NS, not significant.

relative to sinusoids (Fig. 3b), because sinusoids are present throughout the bone marrow.

α -catulin-GFP⁺*c-kit*⁺ HSCs also significantly differed from random spots in their proximity to transition zone blood vessels because they were less likely than random spots to be within 25 μ m of these vessels (Fig. 3c). Only 12 \pm 2.1% of HSCs were within 10 μ m of a transition zone blood vessel (Fig. 3f). As transition zone vessels occupy the outer 20% of bone marrow by volume, the depletion of HSCs near these vessels is consistent with our observation that HSCs are less common near the endosteum (Fig. 2a).

Overall, 85 \pm 3.1% of α -catulin-GFP⁺*c-kit*⁺ HSCs were closer to sinusoids than other blood vessels (significantly more than random spots; Fig. 3g, i). Only 7.7 \pm 2.3% of HSCs were closest to an arteriole (not significantly different from random spots; Fig. 3g, h) and 6.8 \pm 1.8% were closest to a transition zone vessel (significantly fewer than random spots; Fig. 3g, j). We did not detect any differences between male and female mice in HSC localization (Extended Data Fig. 9d–k).

The ability to deep image large segments of bone marrow allowed us to localize much larger numbers of HSCs than previous studies. This allowed us to compare the localization of dividing Ki-67⁺ α -catulin-GFP⁺*c-kit*⁺ HSCs (Fig. 4b, c; which accounted for 15 \pm 2.0% of HSCs) and non-dividing Ki-67⁻ α -catulin-GFP⁺*c-kit*⁺ HSCs (Fig. 4a). Both were most closely associated with sinusoids (Fig. 4e): 81 \pm 5.9% of Ki-67⁻ α -catulin-GFP⁺*c-kit*⁺ cells and 79 \pm 14% Ki-67⁺ α -catulin-GFP⁺*c-kit*⁺ cells were within 10 μ m of a sinusoid (Fig. 4h). Notably fewer dividing and non-dividing HSCs were within 10 μ m of an arteriole (Fig. 4d, g) or a transition zone vessel (Fig. 4f, i). Most of the differences between dividing and non-dividing HSCs were not statistically significant, with the exception that dividing HSCs were significantly more likely to be closer to transition zone vessels.

Overall, 81 \pm 6.0% of Ki-67⁻ α -catulin-GFP⁺*c-kit*⁺ non-dividing HSCs were most closely associated with sinusoids, 9.0 \pm 6.8% with transition zone vessels, and 9.9 \pm 3.7% with arterioles (Fig. 4j). Ki-67⁺ α -catulin-GFP⁺*c-kit*⁺ dividing HSCs were significantly more

likely than $Ki-67^- \alpha\text{-catulin-GFP}^+ c\text{-kit}^+$ non-dividing HSCs to localize close to the bone surface (Fig. 4k).

Based on reconstitution assays (Table 1), at least 30% of $\alpha\text{-catulin-GFP}^+ c\text{-kit}^+$ cells are HSCs. Therefore, some of the $\alpha\text{-catulin-GFP}^+ c\text{-kit}^+$ cells that we imaged are probably not HSCs. If HSCs home to, and engraft, in haematopoietic tissues in competitive transplantation assays with less than perfect efficiency, the HSC purity in this population would be higher than 30%. Even if purity is only 30% and all of the contaminating non-HSCs we imaged were associated with sinusoids, our data would still demonstrate that there are more HSCs associated with sinusoids than arterioles or bone surfaces (for example, subtract 70% from the sinusoid bar in Fig. 3g).

Our results are not consistent with the idea that most quiescent HSCs reside in arteriolar niches associated with NG2-CreER-expressing stromal cells²⁴. While a previous study²⁴ concluded that NG2⁺ Nestin^{high} cells, not Lepr⁺ cells, express the highest levels of *Scf* and *Cxcl12*, the RNA-seq data on which this conclusion was based showed that the Nestin^{high}Lepr⁺ cells analysed in this study were negative for *Nes* and positive for *Lepr* expression (see GSE48764 in the Gene Expression Omnibus)²⁴. Thus, these data are consistent with our data in showing that the cells that express *Scf* and *Cxcl12* are Lepr⁺ (ref. 10).

To address this issue directly, we generated NG2-creER; *Rosa^{tdTomato/+};Scf^{GFP/+}* and NG2-creER; *Rosa^{YFP/+};Cxcl12^{DsRed/+}* mice (YFP, yellow fluorescent protein). While 97% of *Scf-GFP*⁺ stromal cells and 96% of *Cxcl12-DsRed*^{high} stromal cells were Lepr⁺, we did not detect any recombination by NG2-CreER in these cells (Extended Data Fig. 10a, b, g, h). We also conditionally deleted *Scf* or *Cxcl12* with NG2-CreER but did not detect any effect on bone marrow cellularity, HSC frequency, haematopoietic progenitor frequency, or bone marrow reconstituting capacity upon transplantation into irradiated mice (Extended Data Fig. 10c–f, i–l). NG2-CreER-expressing cells are therefore not a source of SCF or Cxcl12 for HSC maintenance in the bone marrow.

Our data provide little support for the idea that dividing and non-dividing HSCs reside in spatially distinct niches, with the exception that dividing HSCs were more likely than non-dividing HSCs to localize near the endosteum. Nonetheless, it remains possible that there are distinct perisinusoidal domains for dividing and non-dividing HSCs.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 January 2015; accepted 27 July 2015.

Published online 23 September 2015.

- Morrison, S. J. & Scadden, D. T. The bone marrow niche for haematopoietic stem cells. *Nature* **505**, 327–334 (2014).
- Kiel, M. J., Radice, G. L. & Morrison, S. J. Lack of evidence that haematopoietic stem cells depend on N-cadherin-mediated adhesion to osteoblasts for their maintenance. *Cell Stem Cell* **1**, 204–217 (2007).
- Kiel, M. J., Yilmaz, O. H., Iwashita, T., Terhorst, C. & Morrison, S. J. SLAM family receptors distinguish haematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109–1121 (2005).
- Lo Celso, C. *et al.* Live-animal tracking of individual haematopoietic stem/progenitor cells in their niche. *Nature* **457**, 92–96 (2009).
- Ding, L. & Morrison, S. J. Haematopoietic stem cells and early lymphoid progenitors occupy distinct bone marrow niches. *Nature* **495**, 231–235 (2013).
- Ding, L., Saunders, T. L., Enikolopov, G. & Morrison, S. J. Endothelial and perivascular cells maintain haematopoietic stem cells. *Nature* **481**, 457–462 (2012).
- Kobayashi, H. *et al.* Angiocrine factors from Akt-activated endothelial cells balance self-renewal and differentiation of haematopoietic stem cells. *Nature Cell Biol.* **12**, 1046–1056 (2010).
- Greenbaum, A. *et al.* CXCL12 in early mesenchymal progenitors is required for haematopoietic stem-cell maintenance. *Nature* **495**, 227–230 (2013).
- Poulos, M. G. *et al.* Endothelial Jagged-1 is necessary for homeostatic and regenerative hematopoiesis. *Cell Rep.* **4**, 1022–1034 (2013).

- Zhou, B. O., Yue, R., Murphy, M. M., Peyer, J. G. & Morrison, S. J. Leptin-receptor-expressing mesenchymal stromal cells represent the main source of bone formed by adult bone marrow. *Cell Stem Cell* **15**, 154–168 (2014).
- Oguro, H., Ding, L. & Morrison, S. J. SLAM family markers resolve functionally distinct subpopulations of haematopoietic stem cells and multipotent progenitors. *Cell Stem Cell* **13**, 102–116 (2013).
- Sugiyama, T., Kohara, H., Noda, M. & Nagasawa, T. Maintenance of the haematopoietic stem cell pool by CXCL12-CXCR4 chemokine signaling in bone marrow stromal cell niches. *Immunity* **25**, 977–988 (2006).
- Omatsu, Y., Seike, M., Sugiyama, T., Kume, T. & Nagasawa, T. Foxc1 is a critical regulator of haematopoietic stem/progenitor cell niche formation. *Nature* **508**, 536–540 (2014).
- Méndez-Ferrer, S. *et al.* Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. *Nature* **466**, 829–834 (2010).
- Kunisaki, Y. *et al.* Arteriolar niches maintain haematopoietic stem cell quiescence. *Nature* **502**, 637–643 (2013).
- Morikawa, S. *et al.* Prospective identification, isolation, and systemic transplantation of multipotent mesenchymal stem cells in murine bone marrow. *J. Exp. Med.* **206**, 2483–2496 (2009).
- Nombela-Arrieta, C. *et al.* Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. *Nature Cell Biol.* **15**, 533–543 (2013).
- Sacchetti, B. *et al.* Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic microenvironment. *Cell* **131**, 324–336 (2007).
- Zhou, B. O., Ding, L. & Morrison, S. J. Hematopoietic stem and progenitor cells regulate the regeneration of their niche by secreting Angiopoietin-1. *eLife* **4**, e05521 (2015).
- Omatsu, Y. *et al.* The essential functions of adipo-osteogenic progenitors as the haematopoietic stem and progenitor cell niche. *Immunity* **33**, 387–399 (2010).
- Hooper, A. T. *et al.* Engraftment and reconstitution of hematopoiesis is dependent on VEGFR2-mediated regeneration of sinusoidal endothelial cells. *Cell Stem Cell* **4**, 263–274 (2009).
- Parmar, K., Mauch, P., Vergilio, J. A., Sackstein, R. & Down, J. D. Distribution of haematopoietic stem cells in the bone marrow according to regional hypoxia. *Proc. Natl Acad. Sci. USA* **104**, 5431–5436 (2007).
- Spencer, J. A. *et al.* Direct measurement of local oxygen concentration in the bone marrow of live animals. *Nature* **508**, 269–273 (2014).
- Kunisaki, Y. *et al.* Arteriolar niches maintain haematopoietic stem cell quiescence. *Nature* **502**, 637–643 (2013).
- Dodd, H. U. *et al.* Ultramicroscopy: three-dimensional visualization of neuronal networks in the whole mouse brain. *Nature Methods* **4**, 331–336 (2007).
- Becker, K., Jahrling, N., Saghati, S., Weiler, R. & Dodd, H. U. Chemical clearing and dehydration of GFP expressing mouse brains. *PLoS ONE* **7**, e33916 (2012).
- Yokomizo, T. *et al.* Whole-mount three-dimensional imaging of internally localized immunostained cells within mouse embryos. *Nature Protocols* **7**, 421–431 (2012).
- Janssens, B., Staes, K. & van Roy, F. Human alpha-catulin, a novel alpha-catenin-like molecule with conserved genomic structure, but deviating alternative splicing. *Biochim. Biophys. Acta* **1447**, 341–347 (1999).
- Yamazaki, S. *et al.* Nonmyelinating Schwann cells maintain haematopoietic stem cell hibernation in the bone marrow niche. *Cell* **147**, 1146–1158 (2011).
- Li, X. M., Hu, Z., Jorgenson, M. L. & Slayton, W. B. High levels of acetylated low-density lipoprotein uptake and low tyrosine kinase with immunoglobulin and epidermal growth factor homology domains-2 (Tie2) promoter activity distinguish sinusoids from other vessel types in murine bone marrow. *Circulation* **120**, 1910–1918 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements S.J.M. is a Howard Hughes Medical Institute investigator, the Mary McDermott Cook Chair in Pediatric Genetics, the director of the Hamon Laboratory for Stem Cells and Cancer, and a Cancer Prevention and Research Institute of Texas Scholar. J.G.P. is a National Science Foundation Graduate Research Fellow. M.M.M. was supported by a National Research Service Award from NIH. This work was supported by the NIH NHLBI (HL097760) and NIH Shared Instrumentation grant NIH S10RR029731. We thank K. Correll and M. Gross for mouse colony management; N. Loof and the Moody Foundation Flow Cytometry Facility; and A. Bugde of the University of Texas Southwestern Live Cell Imaging Facility; and Y. Liu from the Baylor College of Dentistry microCT facility. We also gratefully acknowledge D. Miranda and A. Tully from Bitplane, S. Terclavers from Zeiss, and L. Smith and H. Pudavar from Leica.

Author Contributions M.A., K.S.K., M.M.M., J.G.P., and S.J.M. conceived various aspects of the project, designed, and interpreted experiments. M.A. found $\alpha\text{-catulin}$ is highly restricted in expression to HSCs, and made and characterized the $\alpha\text{-catulin}^{\text{GFP/+}}$ mice. Experiments were performed by M.A., K.S.K., M.M.M., J.G.P., C.N.I., and H.O. with technical assistance from C.J. The confocal imaging and 3D rendering protocols were developed by M.A., K.S.K., M.M.M., and K.L.P. Z.Z. and J.G.P. performed computational image analysis. The manuscript was written by M.A., K.S.K., M.M.M., J.G.P., Z.Z. and S.J.M.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.J.M. (sean.morrison@utsouthwestern.edu).

METHODS

Mice. The targeting construct for α -catulin^{GFP/+} mice was generated by recombination³¹. Linearized targeting vector was electroporated into Bruce4 ES cells. Correctly targeted ES cell clones were identified by Southern blotting and injected into C57Bl/6-Tyr^{c-2j} blastocysts. The resulting chimaeric mice were bred with C57Bl/6-Tyr^{c-2j} mice to obtain germline transmission. Then the *Frt-Neo-Frt* cassette introduced by the targeting vector was removed by mating with FLPe mice³². These mice were backcrossed onto a C57Bl/Ka background and germline transmission was checked by PCR. C57Bl/Ka-Thy-1.1(CD45.2) and C57Bl/Ka-Thy-1.2(CD45.1) mice were used in transplant experiments. Male and female mice from 8 to 12 weeks old were used for all studies. *Scf*^{GFP/+} and *Scf*^{fl} mice⁶, *Cxcl12*^{DSRed/+} and *Cxcl12*^{fl} mice⁵, *Lepr-cre* mice³³, *Rosa26-CAG-loxp-stop-loxp-tdTomato* conditional reporter mice³⁴, *Rosa26-loxp-stop-loxp-EYFP* conditional reporter mice³⁵, and *NG2-CreER* mice³⁶ were all previously described. All mice were housed in AAALAC-accredited, specific-pathogen-free animal care facilities at the University of Texas Southwestern Medical Center. All procedures were approved by the University of Texas Southwestern Institutional Animal Care and Use Committee.

HSC isolation and flow cytometry. Bone marrow cells were isolated by either flushing the long bones (tibias and femurs), or by crushing the bones using a mortar and pestle in Ca²⁺- and Mg²⁺-free Hank's balanced salt solution (HBSS, Gibco) supplemented with 2% heat inactivated bovine serum (Gibco). Spleen cells were prepared by crushing the spleen between two glass slides. The cells were gently passed through a 25-gauge needle then filtered using a 100 µm mesh to generate a single cell suspension. Viable cell number was calculated using a Vi-Cell cell counter (Beckman Coulter) or by counting manually with a haemocytometer. For HSC identification by flow cytometry, the cells were stained with antibodies against CD150 (TC15-12F12.2), CD48 (HM48-1), Sca1 (E13-161-7), and c-kit (2B8), as well as the following lineage markers: CD42d (1C2), CD2 (RM2-5), CD3 (17A2), CD5 (53-7.3), CD8 (53-6.7), B220 (6B2), Ter119 (TER-119), and Gr1 (8C5). Antibody staining of cell suspensions was always performed at 4 °C for 20 min. After antibody staining, the cells were stained with the viability dyes 4',6-diamidino-2-phenylindole (DAPI, 2 µg ml⁻¹ in PBS) or propidium iodide (PI, 1 µg ml⁻¹) to exclude dead cells during flow cytometry. To identify other haematopoietic progenitors (common lymphoid progenitors, common myeloid progenitors, granulocyte macrophage progenitors and megakaryocyte and erythroid progenitors; see Extended Data Fig. 3), we stained cell suspensions with antibodies against CD34 (RAM34), CD127 (IL7R α , A7R34), CD16/32 (Fc γ R, 93), Sca1 (E13-161-7), c-kit (2B8) and the lineage markers listed above. Stains that involved CD34-specific antibody were conducted for 90 min on ice. To identify myeloid cells, erythrocytes, megakaryocytes, T cells, B-cell progenitors and B cells, the following antibodies were used: anti-Gr1 (8C5), anti-CD11b (Mac1, M1-70), anti-Ter119 (TER-119), anti-CD41 (MWReg30), anti-CD3 (17A2), anti-CD4 (GK1.5), anti-CD8 (53-6.7), anti-B220 (6B2), anti-IgM (II/41), anti-CD24 (M1/69), and anti-CD43 (1B11). For Extended Data Fig. 10, antibodies against Ter119 (TER-119), CD45 (30-F11), and CD31 (MEC13.3) were used to mark erythrocytes, nucleated haematopoietic cells, and endothelial cells respectively. Goat anti-mouse leptin receptor biotinylated antibody (BAF497, Fisher Scientific) and Streptavidin-BV421 (405226, Biolegend) were used to mark *Lepr*⁺ cells. Antibodies were conjugated to one of the following dyes depending on the experiment and cell population: PE-Cy5, PerCP-eFluor710, PE-Cy7, PE, APC, APCeFluor 780, APC-H7, PerCP-Cy5.5, eFluor 660, Alexa Fluor 700 and PE-Cy5.

Colony formation in methylcellulose. Individual HSCs were sorted into methylcellulose culture medium (M3434, StemCell Technologies) in 96 well plates (1 cell per well). After sorting, the plates were kept at 37 °C in a cell culture incubator with 6.5% CO₂ and constant humidity for 14 days. Then colonies were counted and identified based on size and morphology using an Olympus IX81 inverted microscope.

Bone marrow preparation from metaphysis and diaphysis for FACS analysis.

To compare bone marrow from the epiphysis/metaphysis to the diaphysis, the metaphysis was separated from the diaphysis using scissors at the point where the central sinus branches (see Extended Data Fig. 7). Then each segment of bone was crushed using a mortar and pestle and small bone fragments were suspended in staining medium (Ca²⁺- and Mg²⁺-free Hank's balanced salt solution (Gibco) supplemented with 2% heat inactivated bovine serum (Gibco) and gently triturated until no marrow was visibly attached to the bone. The cell suspension was filtered through a 100 µm mesh to obtain a single-cell suspension. The cell suspensions were then analysed to determine cellularity and HSC frequency.

Competitive reconstitution assays in irradiated mice. Adult recipient mice were administered a minimum lethal dose of radiation using an XRAD 320 X-ray irradiator (Precision X-Ray) to deliver two doses of 540 rad at least 3 h apart. Cells were transplanted by injection into the retro-orbital venous sinus of anaesthetized recipient mice. 300,000 recipient whole bone marrow (WBM) cells were

transplanted along with the donor cells. In Table 1, HSC frequency was calculated using Extreme Limiting Dilution Analysis³⁷ software (<http://bioinf.wehi.edu.au/software/elda/>) (2–4 independent experiments per cell population). For secondary transplants, 3 million WBM cells from primary recipient mice were transplanted into irradiated secondary recipient mice. Blood was collected from the tail vein of recipient mice at 4-week intervals after transplantation for at least 16 weeks after transplantation. For analysis of the levels of donor cells in peripheral blood, red blood cells were lysed with ammonium potassium buffer, and the remaining cells were stained with antibodies against CD45.1 (A20), CD45.2 (104), CD11b (Mac1, M1-70), Gr-1 (8C5), B220 (6B2), and CD3 (17A2).

Cell cycle analysis. For analysis of DNA content in HSCs and other haematopoietic cells, the cells were isolated by flow cytometry as described above and sorted directly into 70% ethanol then stored at –20 °C for at least 24 h. The cells were washed multiple times with staining medium (see above) then incubated in staining medium containing 50 µg ml⁻¹ propidium iodide (Molecular Probes) for 30 min at room temperature and analysed using a FACSaria or FACSCanto II flow cytometer (BD Biosciences). Data were analysed using FACSDiva (BD Biosciences) or FlowJo (Tree Star) software. To assess 5-bromo-2'-deoxyuridine (BrdU) incorporation *in vivo*, mice were intraperitoneally injected with a single dose of BrdU (1 mg BrdU per 10 g body mass) then maintained on 0.5 mg BrdU per ml of drinking water for 3 days. For analysis of BrdU incorporation into HSCs, bone marrow cells were stained with the following antibodies that were selected to survive fixation: anti-CD150-BV421, anti-CD48-AF700, anti-CD2-PE, anti-CD3-PE, anti-CD5-PE, anti-CD8-PE, anti-Ter119-PE, anti-Gr1-PE, anti-Sca1-PerCP-Cy5.5 and c-kit-APCH7 (BD Biosciences; antibody clones are as described above for HSC isolation and flow cytometry). For isolation of α -catulin-GFP⁺-c-kit⁺ cells, bone marrow cells were stained with anti-c-kit-APCH7 antibody. After antibody staining, the target cell populations were double sorted to ensure purity, then fixed and stained with an anti-BrdU antibody using the BrdU APC Flow Kit (BD Biosciences) according to the manufacturer's instructions.

Sorting of α -catulin-GFP⁺-c-kit⁺ cells to determine cell diameter. Bone marrow cells from α -catulin-GFP⁺ mice were prepared for flow cytometric analysis as described above. Biotinylated anti-c-kit antibody (2B8 clone, eBiosciences) followed by streptavidin-AF647 (Life Technologies) were used to stain bone marrow cells. α -catulin-GFP⁺-c-kit⁺ cells were sorted into a drop of staining medium on a slide coated with poly-D-lysine (0.5 mg ml⁻¹ poly-D-lysine in water was used to coat the slides over night at room temperature). Slides were incubated for 45 min at 4 °C to let the sorted cells attach to the slide surface. A 16% paraformaldehyde (PFA) stock solution was then added gently into the drop of staining medium to achieve a final PFA concentration around 4%. Cells were fixed for 10 min at room temperature and washed multiple times with PBS. The cells were then stained with DAPI (2 µg ml⁻¹ in PBS), with 0.1% IgePal630 (Sigma) for 30 min, followed by multiple washes in PBS. ProLong Gold antifade (Life Technologies) was used to mount the cells. An LSM780 confocal microscope (Zeiss) was used to image the cells and Imaris software was used to measure cell diameter.

MicroCT analysis of bones. Dissected intact femurs from 10–12 week old littermate mice were fixed in 4% PFA overnight at 4 °C. The bones were washed multiple times with and stored in 70% ethanol until they were scanned using a Scanco Medical microCT 35 machine at the Texas A&M University Baylor College of Dentistry. The scan was performed with a 3.5 µm voxel size resolution, 55 kV, 145 µA, and an integration time of 800 ms. Scanco software was used for analysis. A common reference point was determined for all bones scanned based on the growth plate, and trabecular and cortical regions were analysed for each bone.

PCR genotyping. The following primers were used to genotype α -catulin^{GFP} allele: Cin-G1, 5'-GAAGTAGTGGCACAAGGGTAGGGG-3'; Cin-G2, 5'-GGCCGCGTACCTGAGAAAC-3'; Cin-G3, 5'-GTTGCCGTCGTCCTTG AAGAAAG-3'. Genotyping primers for *Cxcl12*^{DSRed} mice⁵ were previously reported.

Half bone whole-mount tissue preparation for imaging. Freshly dissected tibias from 8–12 week old mice were fixed in cold 4% PFA in PBS (Affymetrix) for 7–8 h at 4 °C while shaking. The bones were washed with PBS to remove PFA and cryoprotected in 30% sucrose PBS solution overnight at 4 °C with shaking. The bones were embedded in OCT (Fisher) and flash frozen in liquid nitrogen. A Leica cryostat was used to longitudinally bisect the bones. Intact half bone was washed in PBS to remove OCT then processed for staining, clearing and imaging as below.

Bone marrow whole-mount plug preparation for imaging. Intact bone marrow plugs from freshly dissected tibias of 8–12 week old mice were extruded from the bone using a PFA-filled syringe with a 25-gauge needle and placed directly into 4% PFA solution for 3 h at room temperature. Fixed plugs were then washed in PBS before being stained, cleared and imaged as below.

Whole-mount immunostaining. All staining procedures were performed in eppendorf tubes on a rotator at room temperature. The staining solution

contained 10% dimethyl sulfoxide, 0.5% IgePal630 (Sigma), and 5% donkey serum (Jackson Immuno) in PBS. Half bones and plugs were blocked in staining solution containing anti-CD16/32 mouse Fc-blocking antibody (BD Biosciences) and 1% BlokhenII (Aves Labs) overnight at room temperature. After blocking, half bones were stained for three days at room temperature with primary antibodies in staining solution. Then the tissues were washed multiple times in PBS at room temperature for one day and put into staining solution containing secondary antibodies for three days followed by a one-day wash to remove any unbound secondary antibodies. Antibodies used for whole mount staining included chicken anti-GFP (GFP-1020, Aves Labs), goat anti-c-kit (BAF1356, R&D Systems), rabbit anti-laminin (ab7463, abcam), rat anti-Ki-67 (SolA15, eBioscience), Alexa Fluor 647-AffiniPure F(ab')₂ fragment donkey anti-chicken IgY, Alexa Fluor 488-AffiniPure F(ab')₂ fragment donkey anti-rabbit IgG, AMCA-AffiniPure F(ab')₂ fragment donkey anti-rabbit Ig, Alexa Fluor 488-AffiniPure F(ab')₂ fragment donkey anti-rat IgG (all from Jackson ImmunoResearch), and 555 or 488 conjugated donkey anti-goat (A-11055 and A-21432 from Life Technologies). For isotype control staining in Extended Data Fig. 8b, goat IgG control (BAF108, R&D Systems), rabbit IgG control (ab27478-100, Abcam), rat IgG control (012-000-003, Jackson Immuno) and non-immune chicken IgY control (N-1010, Aves Labs) were used along with the secondary antibodies described above. The fixation time of the tissue, using 0.5% IgePal630 and 10% DMSO in the staining solution, and incubation of the tissue for 3 days in both primary and secondary antibodies were critical factors for efficient deep penetration of antibodies throughout the whole-mount tissue. For anti-Ki-67 antibody penetration, before the blocking step, treatment of the tissues with 0.05% SDS overnight in PBS was necessary.

Comparison of clearing protocols. Scattering and spherical aberration due to refractive index mismatch limit the maximum depth of penetration of visible light into aqueous tissue to about 100 μm ³⁸. Optical clearing agents can decrease the amount of scattered light and therefore increases the depth of penetration. Most optical clearing agents work by replacing the low refractive index aqueous components of the tissue with agents of higher refractive index to match that of the tissue, such as collagen and cell components. Because each tissue has unique properties, such as density of cells and extracellular matrix, the optimum tissue clearing method must be determined empirically. Bone marrow does not have high lipid content and therefore, unlike brain, is not limited by the opacity of lipids. Therefore clearing methods that remove the lipid with SDS by either electrophoresis³⁹ or passive flow⁴⁰ were ineffective and had the additional disadvantage of SDS destruction of cell surface epitopes. Bone marrow is very densely packed with cells, which probably explains why optical clearing agents and methods with lower refractive indices such as Sca/eA2 (ref. 41), CUBIC⁴², and Focus Clear³⁹ did not efficiently clear the marrow in our hands. We found that Murray's clear (1:2 Benzyl Alcohol: Benzyl Benzoate; BABB)⁴³ was the most effective clearing method, but is only compatible with antibodies conjugated with stable chemical fluorophores, such as the Alexa Fluor dyes. Murray's clear rapidly quenches fluorescent proteins, so GFP⁺ cells were identified using an antibody against GFP when performing deep imaging of tissues. For imaging bone marrow cells from *Cxcl12^{DsRed}* or *Lepr-cre;tdTomato* mice we used the 3DISCO⁴⁴ clearing method because tetrahydrofuran and dibenzyl ether preserve endogenous fluorescence better than BABB, although they did not clear the tissue as effectively as Murray's clear.

Tissue clearing using modified Murray's clear. All clearing of half bones and bone marrow plugs was performed in eppendorf tubes on a rotator at room temperature. Immunostained tissues were washed in PBS and dehydrated in either a methanol or ethanol dehydration series then incubated for 3 h in methanol or overnight in ethanol with several changes of 100% alcohol. The alcohol was then exchanged with BABB. The tissues were incubated in BABB for 3 h to overnight with several exchanges of fresh BABB. Half bones or bone marrow plugs were mounted in BABB between two coverslips and sealed with silicone (premium waterproof silicone II clear, General Electric). As previously published²⁶, we found it necessary to clean the BABB of peroxides (which can accumulate as a result of exposure to air and light) by adding 10 g of activated aluminium oxide (Sigma) to 40 ml of BABB and rotating for at least 1 h, then centrifuging at 2000g for 10 min to remove the suspended aluminium oxide particles.

Confocal imaging of thick tissue. Three-dimensional (3D) confocal microscopy of the bone marrow at submicron resolution required specialized equipment. We used both a Zeiss LSM780 and a Leica SP8 resonant scanning confocal. Specifications for the Zeiss LSM780: AxioExaminer upright stand; 405, 488, 561, 594 and 633 nm visible laser lines; internal 32-channel GaAsP detector; Prior OptiScan motorized stage; Coherent Chameleon Vision II pulsed NIR laser for two photon excitation; Zeiss BiG two channel nondescanned detector. Specifications for the Leica: Acousto Optical Beam Splitter, Spectral detection, 8 kHz Resonant tandem scanner, HyD hybrid detectors, and 405, 488, 561, 633 nm laser lines. The optimum clearing agent (BABB) for bone marrow has a

refractive index of 1.56, similar but not identical to standard immersion oil. Deep imaging also requires a long working distance objective. For the Zeiss LSM780 we found the best available objective was a Zeiss LD LCI Plan-Apo 25 \times /0.8 multi-immersion objective lens, which has a 570 μm working distance. We used Immersol 518F immersion oil for Zeiss LSM780 imaging. For the Leica SP8, we used an HCX APO L20 \times /0.95 BABB immersion objective with a 1.95 mm working distance. High resolution imaging of large volumes of thick tissue by acquisition of tiled Z-stacks is very time consuming, thus it was important to optimize the acquisition settings for each microscope to minimize acquisition time while preserving adequate resolution and signal to noise ratio. On the Zeiss LSM780, images were taken at 512 \times 512 pixel resolution with 2 μm Z-steps, pinhole for the internal detector at 47.7 μm . Bone was imaged by second harmonic generation (SHG) with 850 nm pulsed near infrared (NIR) excitation using the non-descanned detector. On the Leica SP8, images were taken using the resonance scanner using 8 \times line averaging with the minimum zoom of 1.25 \times at 812 \times 812 pixel resolution, pinhole at 44.7 μm , and 2 μm Z-steps.

Image annotation and analysis. Confocal tiled Z-stack images were rendered in 3D and analysed using Bitplane Imaris v7.7.1 software installed on a Dell Precision T7610 64-bit workstation with Dual Intel Xeon Processor E5-2687W v2 (Eight Core HT, 3.4GHz Turbo, 25 MB), 128 GB RAM, and 16 GB AMD FirePro W9100 graphics card. Individual α -catulin-GFP⁺c-kit⁺ HSCs were identified using the ortho slicer function in Imaris software to visualize digital serial sections of the large 3D image. We identified HSCs as having a round morphology, with GFP throughout the cell, and c-kit expression surrounding the cell surface. These criteria prevented false positive identification of cellular debris or α -catulin-GFP⁺c-kit⁺ endothelial cells with elongated cell body morphology. HSC coordinates and size were interactively annotated using the Imaris spots function in manual mode. Bone and non-myelinating Schwann cells were segmented based on thresholding of the SHG (which detects collagen fibres in bone) or GFAP channels, respectively, using the Imaris surface function. Cortical and trabecular bone were then divided into separate surfaces interactively based on SHG signal and morphology. We used anti-laminin antibodies to immunofluorescently label all of the vasculature within the bone marrow. Arteries, arterioles, and capillaries have continuous basement membranes, which are observable as uniform laminin staining. In contrast, bone marrow sinusoids have a discontinuous fenestrated basement membrane⁴⁵. Laminin staining of sinusoids clearly demonstrates discontinuous basement membranes, thereby allowing unambiguous identification of sinusoidal vessels in the absence of other markers⁴⁵. Because we were able to image the entire marrow cavity, we were able to trace and digitally label each artery and all of its subsequent branching into smaller arteries and arterioles as they approached the endosteal surface. This is an advantage of the deep imaging approach that allowed us to unambiguously identify vessels in a way that is not possible in thin sections where the connectivity of vessels cannot be traced.

Near the endosteum, arterioles connect to the smallest diameter vessels of the capillary network that line the endosteum. These capillaries then connect to larger diameter sinusoidal vessels. By following the blood vessel paths in six samples, we determined that in the diaphysis, the outer 20% of the marrow by volume contained all of the vessels involved in the transition from arteriole to sinusoid; the most distal portion of the arterioles, the connecting capillaries, and the initial portion of sinusoids. We identified this region as the transition zone in keeping with published criteria³⁰. Therefore, we used the published morphological characteristics of orientation, location, and basement membrane continuity to subdivide blood vessels within the bone marrow. The Imaris surface function was used to create three distinct digital surfaces corresponding to each type of blood vessel. SHG signal was used to create bone surfaces. Three-dimensional distances between HSCs and digital vessel or bone surfaces were calculated using the Imaris Distance Transform Matlab XTension and volumetric decile calculations were performed using a Matlab-based Imaris XTension. The annotated programs, entitled 'Visualizing Progressive Zones of Equal Volume in a 3D Tissue (Matlab Extensions for Imaris)' are available for download from the Morrison lab protocols webpage at the CRI website under 'More Information' (<http://cri.utsu.edu/sean-morrison-laboratory/more-information/>).

Random spot generation and insertion. The original 3D images of the GFP channel in the Imaris format were used to generate random spots for each sample. In a 16 bit Imaris file of the original bone marrow image, 3D voxels were represented by signal intensity values that ranged between 0 and 65,535. Those signal intensity values were imported into MATLAB using the imreadBF package (<http://www.mathworks.com/matlabcentral/fileexchange/32920-imread-for-multiple-life-science-image-file-formats>) and the Bio-Formats software (<http://www.openmicroscopy.org/site/products/bio-formats>). The images were filtered to exclude low-intensity regions that included blood vessel lumens and fat bodies where HSCs were not found and random spots were not generated. The intensity-filtered images, of which the excluded portions were given zero intensity values,

were processed with MATLAB's median filter to remove 'salt-and-pepper noise'. Then the intensity images were converted to binary signals by turning any non-zero intensity value into 'one'. Those 'one' signals were used to determine the voxel locations that were used to generate random spots. The locations were randomly permuted, and enough random spot coordinates were designated to approximate the random spot distribution (more than 50,000 per bone). The random spot coordinates were transferred to Imaris to generate the random spots, and distances to cell types or landmarks in the bone marrow were calculated based on the distance transformation files generated using landmark surfaces such as arterioles, sinusoids, transition zone vessels, and bone. We confirmed the randomness of the distribution of the random spots by measuring the frequency of random spots in each percentile of bone marrow volume. Random spots were given a diameter of 6 μm , similar to the observed average HSC diameter.

Statistical methods. To assess whether the distribution of HSCs significantly differed from random spots with respect to particular bone marrow landmarks, we used a normalized two-sample Kolmogorov–Smirnov test. The two-sample Kolmogorov–Smirnov test calculates and evaluates the maximum difference between the empirical cumulative distribution functions (ECDFs) of two test groups where each group is a vector of continuous values, which in our case were the distances from HSCs or random spots to particular bone marrow cell types or structures. As we had multiple biological samples, we normalized them by the following approaches: (1) HSCs and random spot distances from the same sample were pooled to determine a range of distances, which was then used to generate 100 equal-length bins for each sample, so that each bin represented 1% of the distance range for that sample; (2) for each sample, the number of HSCs or random spots in each bin was determined and normalized to percentages; (3) the average percentage of HSCs or random spots in each bin was calculated across all samples; (4) the averaged binned percentages of HSCs or random spots were used to approximate the probability density functions (PDF), and the ECDFs were calculated based on those approximate PDFs; (5) the two-sample Kolmogorov–Smirnov test in MATLAB was used and slightly modified so that it accepted the two normalized ECDFs as inputs, and the Kolmogorov–Smirnov P values were adjusted using the Bonferroni method to account for multiple comparisons.

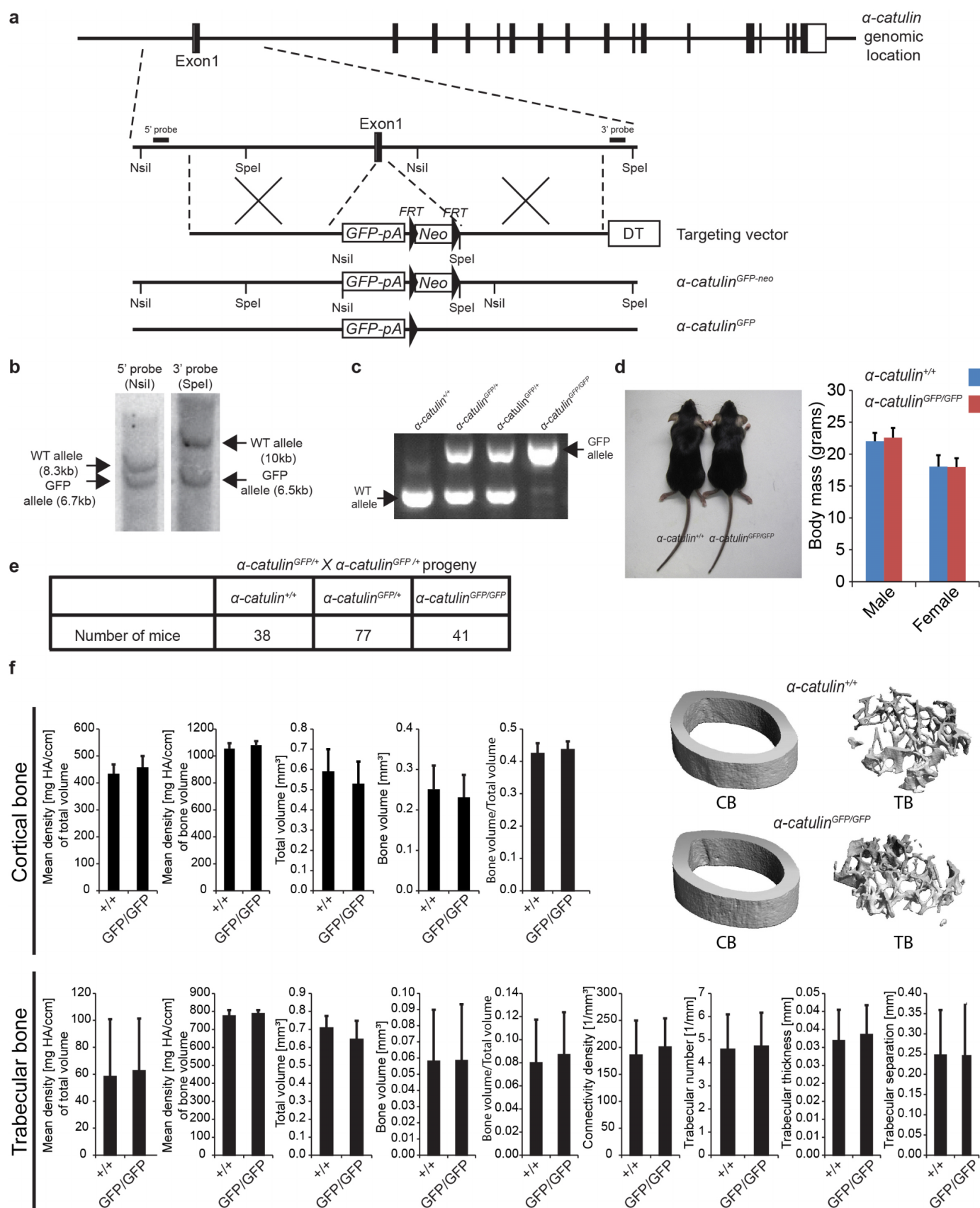
The data presented in the figures reflect multiple independent experiments performed on different days using tissues from different mice. No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded during allocation or outcome assessment. Variation is always indicated using standard deviation. For analysis of the statistical significance of differences between two groups, we first determined that variance in the two groups was similar using an F -test, and then two-tailed Student's t -tests. Single-factor ANOVA and two-way ANOVA with multiple comparisons tests were used for comparisons among three groups in Extended Data Fig. 2.

Not all samples were suitable for image analysis and those that did not meet the criteria were not analysed. Occasionally, the antibody staining was not strong enough for us to detect HSCs or other landmarks in the deepest part of the bone marrow, or the samples were damaged during processing and in these cases the

samples were not analysed. All mice used in our studies were between 8 and 12 weeks old, including both male and female mice. We did not observe any differences in HSC localization between male and female mice (Extended Data Fig. 9d–k), so the data were combined for purposes of analysis. All conclusions were based on data obtained from at least three independent experiments involving samples obtained from different mice and processed on different days.

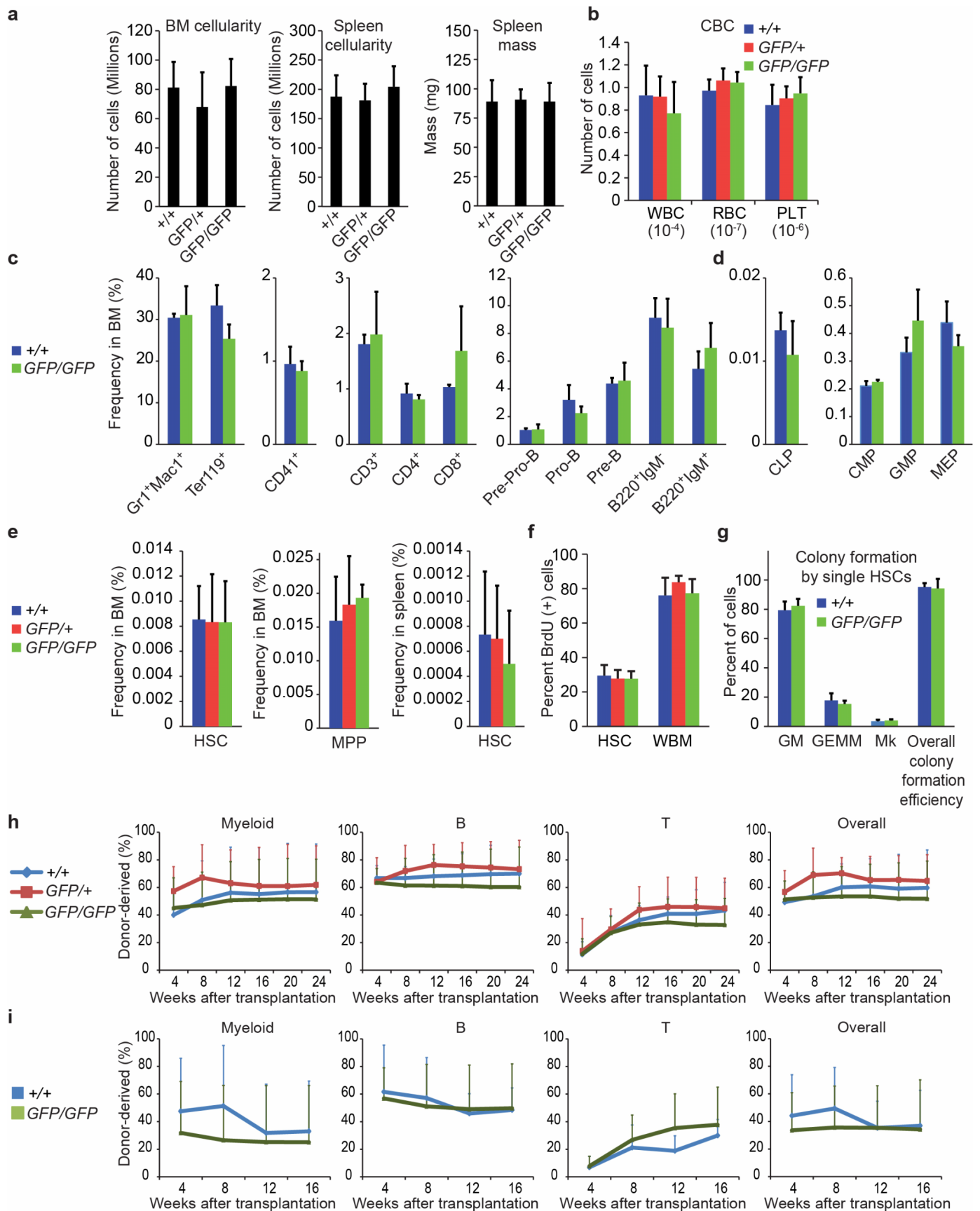
Code availability. Code was written to separate bone marrow into volumetric deciles and to identify the transition zone in the outer 20% of bone marrow. This code is available on the Morrison lab protocols webpage at the CRI website under 'More Information' (<http://cri.uts.edu/sean-morrison-laboratory/more-information/>).

31. Liu, P., Jenkins, N. A. & Copeland, N. G. A highly efficient recombinase-based method for generating conditional knockout mutations. *Genome Res.* **13**, 476–484 (2003).
32. Rodríguez, C. I. *et al.* High-efficiency deleter mice show that *FLPe* is an alternative to *Cre-loxP*. *Nature Genet.* **25**, 139–140 (2000).
33. Mignone, J. L., Kukekov, V., Chiang, A. S., Steindler, D. & Enikolopov, G. Neural stem and progenitor cells in nestin-GFP transgenic mice. *J. Comp. Neurol.* **469**, 311–324 (2004).
34. Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nature Neurosci.* **13**, 133–140 (2010).
35. Srinivas, S. *et al.* Cre reporter strains produced by targeted insertion of *EYFP* and *ECFP* into the *ROSA26* locus. *BMC Dev. Biol.* **1**, 4 (2001).
36. Zhu, X. *et al.* Age-dependent fate and lineage restriction of single NG2 cells. *Development* **138**, 745–753 (2011).
37. Hu, Y. & Smyth, G. K. ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J. Immunol. Methods* **347**, 70–78 (2009).
38. Zhu, D., Larin, K. V., Luo, Q. & Tuchin, V. V. Recent progress in tissue optical clearing. *Laser Photon Rev.* **7**, 732–757 (2013).
39. Chung, K. *et al.* Structural and molecular interrogation of intact biological systems. *Nature* **497**, 332–337 (2013).
40. Yang, B. *et al.* Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell* **158**, 945–958 (2014).
41. Hama, H. *et al.* Scale: a chemical approach for fluorescence imaging and reconstruction of transparent mouse brain. *Nature Neurosci.* **14**, 1481–1488 (2011).
42. Susaki, E. A. *et al.* Whole-brain imaging with single-cell resolution using chemical cocktails and computational analysis. *Cell* **157**, 726–739 (2014).
43. Becker, K., Jahrling, N., Saghaei, S. & Dödt, H. U. Immunostaining, dehydration, and clearing of mouse embryos for ultramicroscopy. *Cold Spring Harb. Protoc.* **2013**, 743–744 (2013).
44. Ertürk, A. *et al.* Three-dimensional imaging of solvent-cleared organs using 3DISCO. *Nature Protocols* **7**, 1983–1995 (2012).
45. Inoue, S. & Osmond, D. G. Basement membrane of mouse bone marrow sinusoids shows distinctive structure and proteoglycan composition: a high resolution ultrastructural study. *Anat. Rec.* **264**, 294–304 (2001).
46. Draenert, K. & Draenert, Y. The vascular system of bone marrow. *Scan Electron Microsc.* **4**, 113–122 (1980).
47. Kopp, H. G., Hooper, A. T., Avecilla, S. T. & Rafii, S. Functional heterogeneity of the bone marrow vascular niche. *Ann. NY Acad. Sci.* **1176**, 47–54 (2009).



Extended Data Figure 1 | Generation of α -catulin^{GFP} mice. **a**, The targeting strategy to generate the α -catulin^{GFP} allele is shown. The targeting vector was generated by retrieving a genomic fragment of the α -catulin gene, including exon 1, from bacterial artificial chromosome clone RP24-146F11 by recombineering³¹. The retrieved genomic region was then modified to replace most of the exon 1 coding region and the exon 1 to intron 1 junction with an *EGFP-bGH-pA-FRT-neo-FRT* cassette in frame with the first ATG of α -catulin. The final targeting vector was then linearized and electroporated into C57Bl-derived Bruce4 ES cells. **b**, New NsiI and SpeI sites introduced with the *EGFP-bGH-pA-FRT-neo-FRT* cassette were used to screen correctly targeted ES cell clones by Southern blotting for 5' and 3' probes. Correctly targeted ES cells were used to generate chimaeric mice. Upon confirmation of germline transmission by PCR, the α -catulin^{GFP-neo} mice were crossed with FLPe mice³², to remove the neomycin resistance cassette. **c**, PCR genotyping of α -catulin⁺

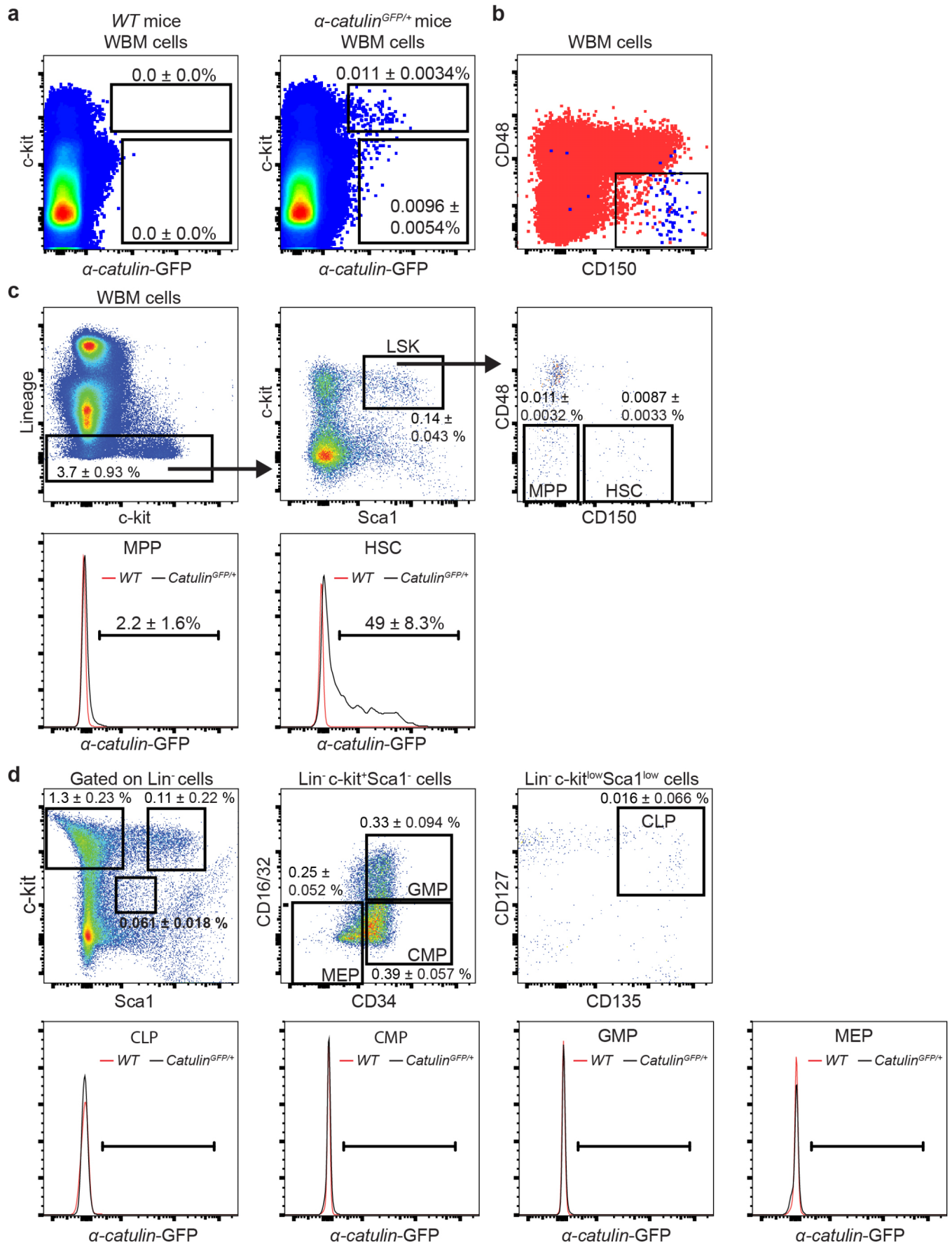
(WT) and α -catulin^{GFP} alleles from α -catulin^{+/+}, α -catulin^{GFP/+}, and α -catulin^{GFP/GFP} mice. **d**, α -catulin^{+/+} and α -catulin^{GFP/GFP} mice did not show any difference in size or body mass ($n = 9$ α -catulin^{+/+} and 8 α -catulin^{GFP/GFP} male mice; $n = 7$ α -catulin^{+/+} and 7 α -catulin^{GFP/GFP} female mice; all were 8–10 weeks old). **e**, α -catulin^{GFP/+} and α -catulin^{GFP/GFP} mice were born at Mendelian frequencies, survived into adulthood in normal numbers, and were apparently developmentally normal. The statistics reflect mice genotypes at 8–10 weeks of age. **f**, Cortical and trabecular femur bone (CB and TB, respectively) did not show any statistically significant differences among α -catulin^{+/+} and α -catulin^{GFP/GFP} mice by microCT (microcomputed tomography) analysis (6 α -catulin^{GFP/GFP} and 5 α -catulin^{+/+} controls at 10–12 weeks of age). HA, hydroxyapatite. All data represent mean \pm s.d. The significance of differences between genotypes was assessed using Student's *t*-tests; none were statistically significant.



Extended Data Figure 2 | α -catulin^{GFP/GFP} mice had normal haematopoiesis, normal HSC frequency, and normal HSC function.

a, Hindlimb bone marrow cellularity ($n = 9$ mice for α -catulin^{+/+}, $n = 4$ mice for α -catulin^{GFP/+} and $n = 9$ mice for α -catulin^{GFP/GFP} genotype), spleen cellularity ($n = 6$ mice for α -catulin^{+/+}, $n = 4$ mice for α -catulin^{GFP/+} and $n = 6$ mice for α -catulin^{GFP/GFP} genotype), and spleen mass ($n = 7$ mice for α -catulin^{+/+}, $n = 4$ mice for α -catulin^{GFP/+} and $n = 7$ mice for α -catulin^{GFP/GFP} genotype). **b**, White blood cell (WBC), red blood cell (RBC), and platelet (PLT) counts per microlitre of peripheral blood from 8–12 week old α -catulin^{+/+}, α -catulin^{GFP/+}, and α -catulin^{GFP/GFP} mice ($n = 9$ mice per genotype). **c, d**, Frequencies of mature haematopoietic cells and progenitors in the bone marrow of 8–12 week old α -catulin^{+/+} and α -catulin^{GFP/GFP} mice (pre-pro-B cells were B220⁺sIgM⁻CD43⁺CD24⁻; pro-B cells were B220⁺sIgM⁻CD43⁺CD24⁺; pre-B cells were B220⁺sIgM⁻CD43⁻; common lymphoid progenitors (CLP) were Lin⁻c-kit^{low}Sca-1^{low}CD127⁺CD135⁺; common myeloid progenitors (CMP) were Lin⁻c-kit⁺Sca-1⁻CD34⁺CD16/32⁻; granulocyte macrophage progenitors (GMP) were Lin⁻c-kit⁺Sca-1⁻CD34⁺CD16/32⁺; and megakaryocyte erythroid progenitors (MEP) were Lin⁻c-kit⁺Sca-1⁻CD34⁻CD16/32⁻ ($n = 3$ mice per genotype). **e**, Bone marrow CD150⁺CD48⁻LSK HSC frequency, bone marrow CD150⁻CD48⁻LSK MPP frequency ($n = 12$ mice per genotype in 12

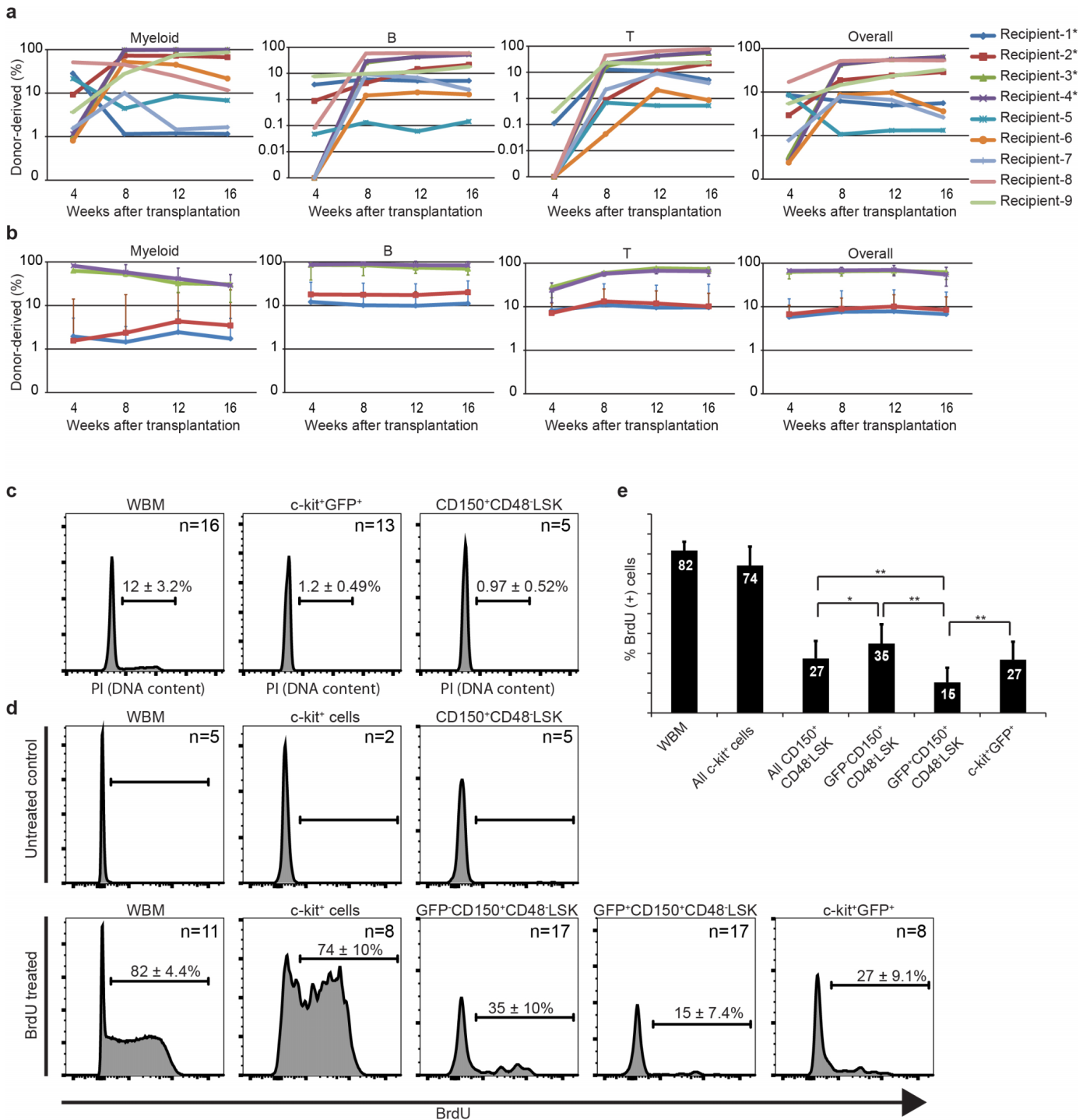
independent experiments), and spleen HSC frequency ($n = 3$ mice per genotype in 3 experiments). **f**, Percentage of HSCs and whole bone marrow cells that incorporated a 3 day pulse of BrdU *in vivo* ($n = 6$ α -catulin^{+/+}, 9 α -catulin^{GFP/+}, and 7 α -catulin^{GFP/GFP} 8–12 week old mice in 3 independent experiments). **g**, Colony formation by HSCs in methylcellulose cultures (GM, granulocyte macrophage colonies; GEMM, granulocyte erythroid macrophage megakaryocyte colonies; Mk, megakaryocyte colonies; $n = 5$ mice per genotype in 5 independent experiments). **h**, Reconstitution of irradiated mice by 300,000 donor bone marrow cells from 8–12 week old α -catulin^{+/+}, α -catulin^{GFP/+}, or α -catulin^{GFP/GFP} mice competed against 300,000 recipient bone marrow cells ($n = 4$ donor mice and 16 recipient mice for α -catulin^{+/+}, $n = 3$ donor mice and 9 recipient mice for α -catulin^{GFP/+}, and $n = 4$ donor mice and 18 recipients for α -catulin^{GFP/GFP} in 3 independent experiments). **i**, Serial transplantation of 3,000,000 WBM cells from primary recipient mice shown in **h** into irradiated secondary recipient mice ($n = 4$ primary α -catulin^{+/+} recipients were transplanted into 17 secondary recipients, and $n = 6$ primary α -catulin^{GFP/GFP} recipients were transplanted into 20 secondary recipients). All data represent mean \pm s.d. The statistical significance of differences between genotypes was assessed using Student's *t*-tests or ANOVAs; none were significant.



Extended Data Figure 3 | α -catulin–GFP expression among

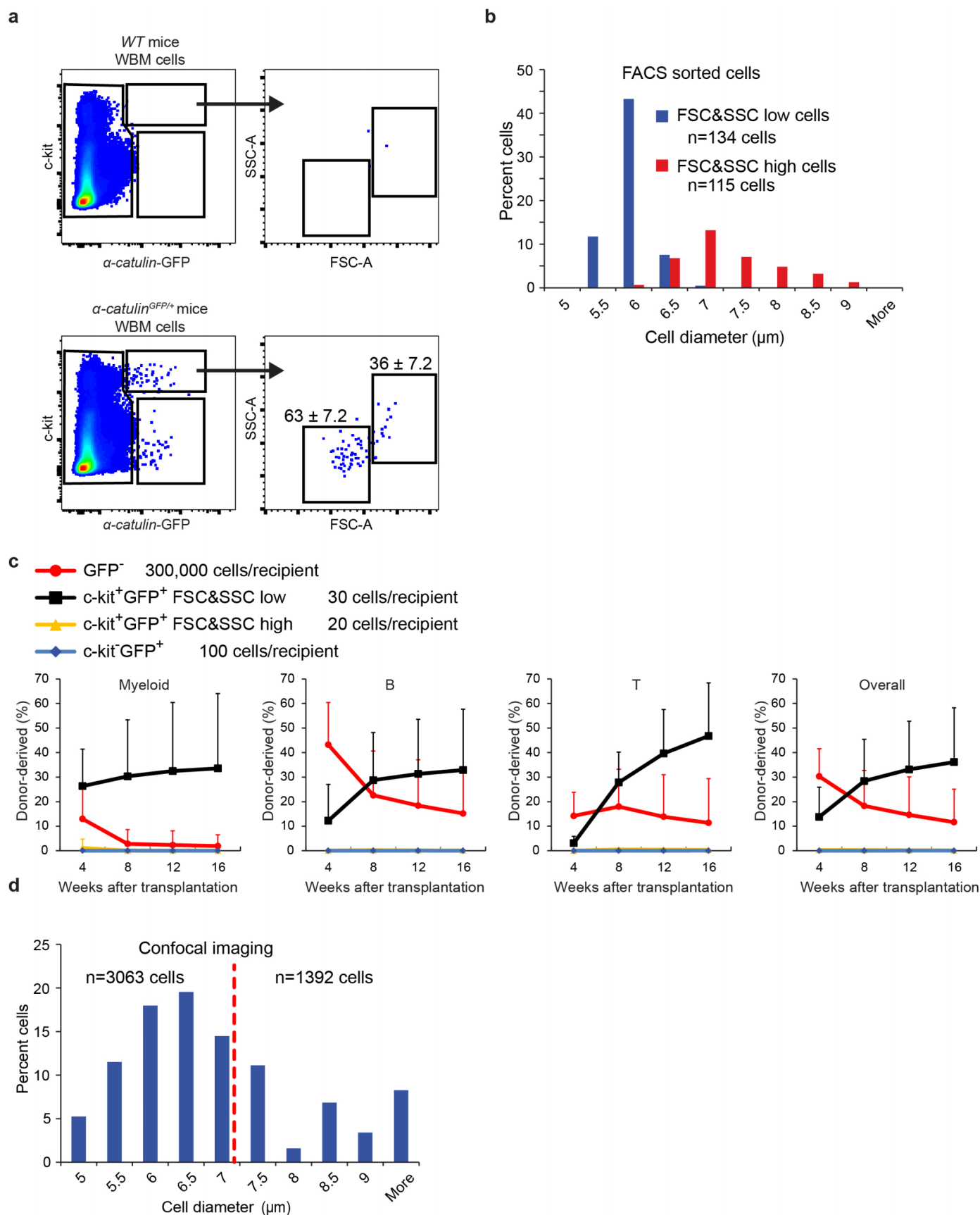
haematopoietic cells is highly restricted to HSCs. **a**, The frequency of α -catulin–GFP⁺ bone marrow cells in negative control α -catulin^{+/+} (WT) mice and α -catulin^{GFP/+} mice ($n = 14$ mice per genotype in 11 independent experiments). In all cases in this figure, percentages refer to the frequency of each population as a percentage of WBM cells. **b**, α -catulin–GFP⁺c-kit⁺ cells from Fig. 1b are shown (blue dots) along with all other bone marrow cells in the same sample (red dots). **c**, CD150⁺CD48[−]LSK HSCs express α -catulin–GFP but CD150[−]CD48[−]LSK MPPs do not ($n = 17$ mice in 12 independent experiments). A minority of the α -catulin–GFP⁺c-kit⁺ cells had high forward scatter, lacked reconstituting potential, and were gated out when isolating HSCs

by flow cytometry and when identifying HSCs during imaging (see Extended Data Fig. 5 for further explanation). **d**, Lin[−]c-kit^{low}Sca-1^{low}CD127⁺CD135⁺ common lymphoid progenitors (CLPs), Lin[−]c-kit⁺Sca-1[−]CD34⁺CD16/32[−] common myeloid progenitors (CMPs), Lin[−]c-kit⁺Sca-1[−]CD34⁺CD16/32⁺ granulocyte-macrophage progenitors (GMPs), and Lin[−]c-kit⁺Sca-1[−]CD34[−]CD16/32[−] megakaryocyte-erythroid progenitors (MEPs) did not express α -catulin–GFP. α -catulin^{GFP/+} and control cell populations had similar levels of background GFP signals that accounted for fewer than 1% of the cells in each population ($n = 9$ mice per genotype in 2 independent experiments).



Extended Data Figure 4 | α -catulin-GFP⁺ c-kit⁺ bone marrow cells are highly enriched for HSC activity and are quiescent. **a**, Competitive reconstitution assays in which 1 donor α -catulin-GFP⁺ c-kit⁺ bone marrow cell was transplanted along with 300,000 recipient bone marrow cells into irradiated recipient mice. Each line represents 1 of the 9 mice (out of 34 transplanted; see Table 1) that were long-term multilineage reconstituted by donor myeloid, B, and T cells. **b**, Three million WBM cells from primary recipient mice 1–4 from **a** (indicated by an asterisk) were transplanted into secondary recipient mice (7 secondary recipients from primary recipient-1; 4 secondary recipients from primary recipient-2; 3 secondary recipients from primary recipient-3; 3 secondary recipients from primary recipient-4 for an overall total of 17 secondary recipients). Each line shows the average (\pm s.d.) levels of donor cell reconstitution in secondary recipient mice from each

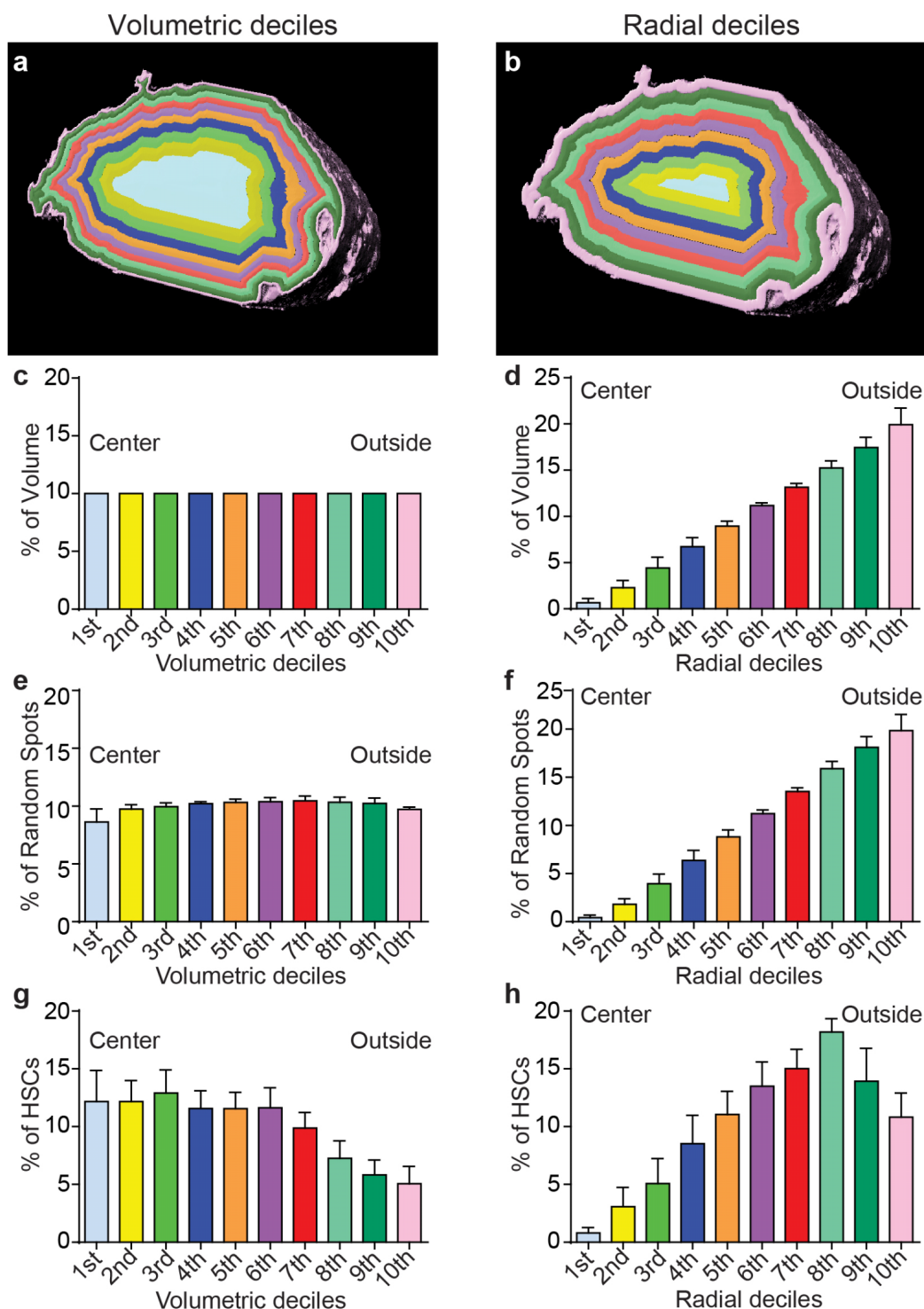
primary donor. **c**, DNA content of WBM cells, α -catulin-GFP⁺ c-kit⁺ HSCs, and CD150⁺ CD48⁻ LSK HSCs. While 11.5% of WBM cells had greater than 2N DNA content (in S/G2/M phases of the cell cycle), only around 1% of α -catulin-GFP⁺ c-kit⁺ HSCs or CD150⁺ CD48⁻ LSK HSCs had greater than 2N DNA content. **d**, BrdU incorporation into WBM cells, c-kit⁺ cells, α -catulin-GFP⁺ CD150⁺ CD48⁻ LSK cells, α -catulin-GFP⁺ CD150⁺ CD48⁻ LSK HSCs, and α -catulin-GFP⁺ c-kit⁺ HSCs after 3 days of continuous BrdU administration (BrdU treated). Untreated negative control mice are also shown. **e**, Percentage of BrdU⁺ cells in each cell population. In each panel, the number of mice used for analysis (without being pooled) is indicated. All data reflect mean \pm s.d. from two to five independent experiments. Statistical significance was assessed using Student's *t*-tests (**P* < 0.05; ***P* < 0.01).



Extended Data Figure 5 | All HSC activity resides among

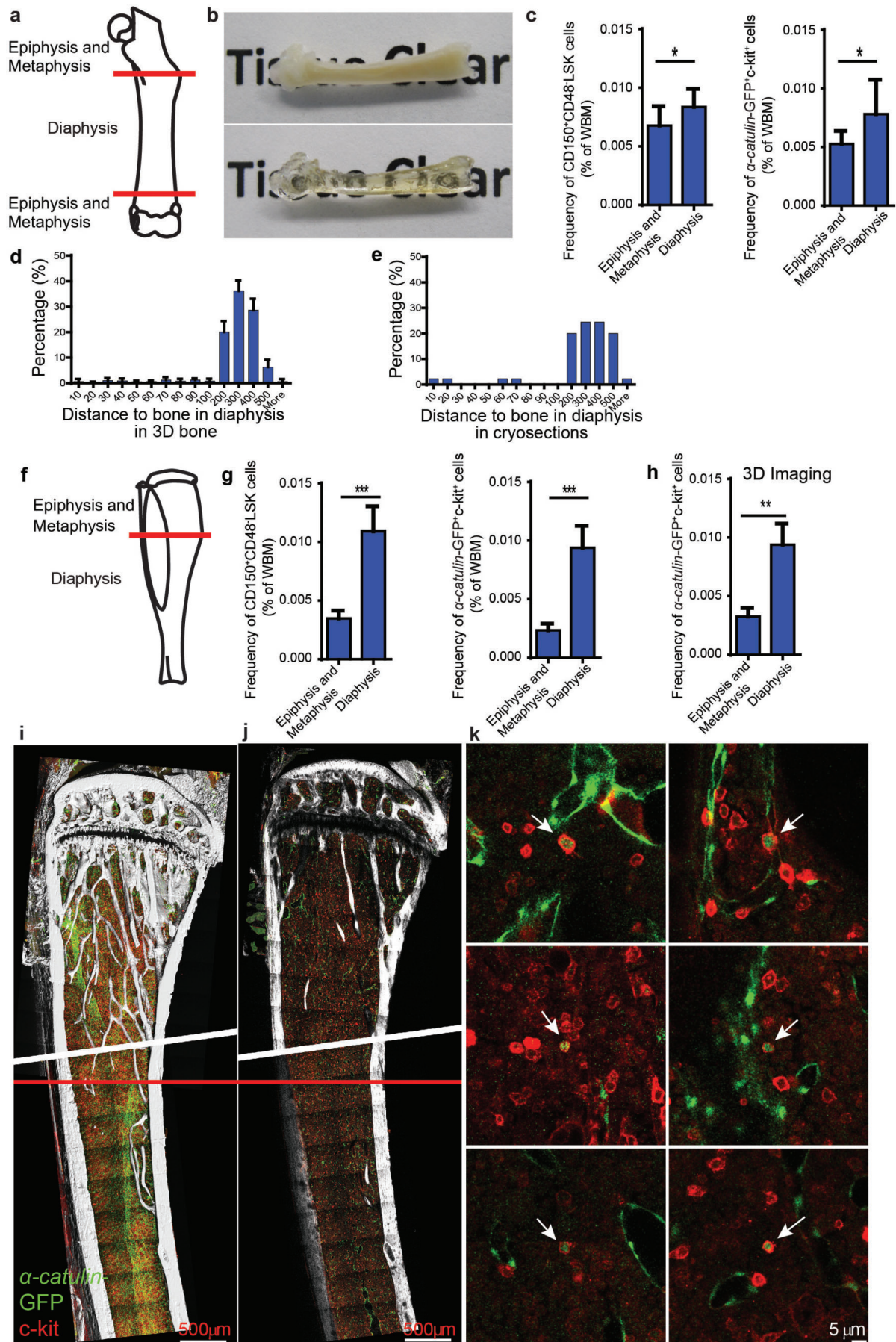
α -catulin⁺GFP⁺c-kit⁺ cells with low forward and side scatter. **a**, Most α -catulin⁺GFP⁺c-kit⁺ cells ($63 \pm 7.2\%$) had low forward and side scatter, but a distinct minority population ($36 \pm 7.2\%$) had higher forward and side scatter that was not typical of HSCs. **b**, We sorted the low scatter and the high scatter α -catulin⁺GFP⁺c-kit⁺ cell populations gated in **a** and measured their diameters (three independent experiments). **c**, Competitive reconstitution assays in irradiated mice revealed that all HSC activity resided in the low scatter cell fraction. For each recipient mouse, the indicated donor cells (based on the number of cells from each population contained within 300,000 bone

marrow cells) were transplanted into irradiated mice along with 300,000 recipient bone marrow cells (mean \pm s.d. from 2 independent experiments with 20 total recipient mice in the GFP⁺ group, 14 total recipient mice in the c-kit⁺GFP⁺ FSC&SSC low group, 11 total recipient mice in the c-kit⁺GFP⁺ FSC&SSC high group, and 9 total recipient mice in the c-kit⁺GFP⁺ group). **d**, The size distribution of all α -catulin⁺GFP⁺c-kit⁺ cells identified by confocal microscopy in bone marrow plugs from the tibia diaphysis (6 bones analysed in 6 independent experiments). In keeping with the flow cytometry data, the largest 40% of imaged cells were not considered HSCs, excluding all cells with diameter larger than 7 μ m.



Extended Data Figure 6 | HSCs are enriched in the central marrow and depleted near the endosteum in the diaphysis. **a, b,** The distribution of HSCs from the central marrow to the endosteum can be determined by drawing concentric cylinders that correspond to equal volumetric deciles from the centre of the marrow to the endosteum (**a**) or to equal radial deciles from the centre to the endosteum (**b**). **c, d,** Each volumetric decile (as in **a**) contains 10% of the marrow volume (**c**). However, cylinders based on radial deciles (as in **b**), contain successively larger volumes of marrow as they approach the endosteum because the circumference of the cylinders becomes larger

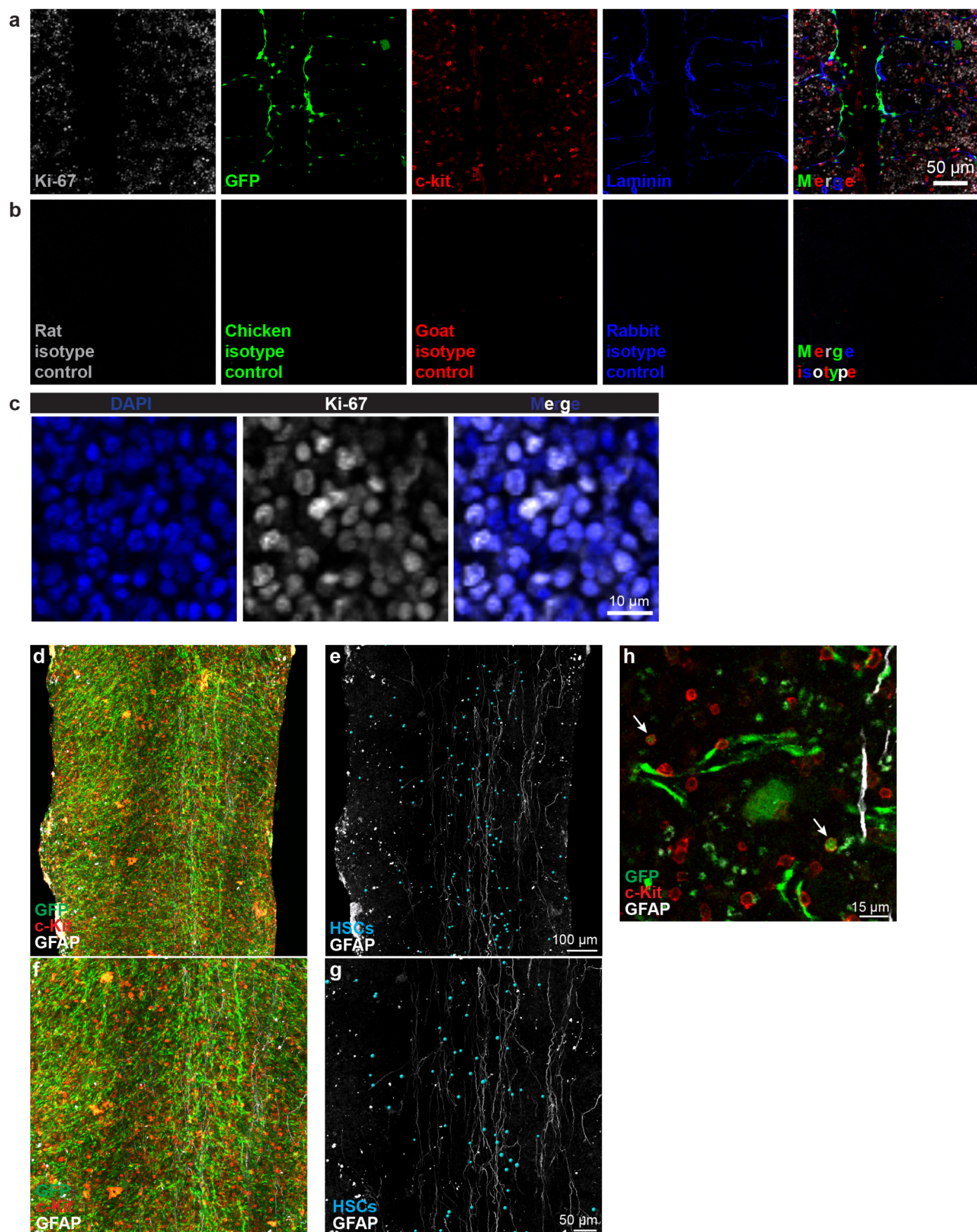
(**d**). **e, f,** The distribution of random spots among volumetric deciles (**a**) is nearly equal because each cylinder contains an equal marrow volume (**e**). However, the number of random spots per cylinder based on radial deciles (**b**) increases from the centre to the endosteum as cylinder volume increases (**f**). **g,** When we plotted our HSC localization data by volumetric deciles (as in Fig. 2a), HSC were enriched towards the central marrow. **h,** When we plotted our HSC localization data by radial deciles, the number of HSCs per cylinder increased towards the endosteum as cylinder volume increased, similar to random spots. All data represent mean \pm s.d.



Extended Data Figure 7 | HSC density is higher in the diaphysis as compared to the metaphysis.

a, Schematic of a femur showing the separation of epiphysis/metaphysis from diaphysis. We divided metaphysis from diaphysis at the point where the central sinus branched (see red line in panels **a**, **f**, and **i**). This is also the point at which the density of trabecular bone declines, moving into the diaphysis. **b**, A bisected femur before and after clearing. **c**, The frequency of CD150⁺CD48⁻LSK cells and α -catulin-GFP⁺c-kit⁺ cells by flow cytometry in the epiphysis/metaphysis versus diaphysis of femurs ($n = 9$ mice in 2 independent experiments). Note that bone marrow cells were extracted from crushed bones. **d**, The distance (μ m) from α -catulin-GFP⁺c-kit⁺ cells to the nearest bone surface in the femur diaphysis based on deep imaging ($n = 368$ cells in 3 bisected femurs). **e**, The distance (μ m) from α -catulin-GFP⁺c-kit⁺ cells to the nearest bone surface in the femur diaphysis based on analysis of thin (7μ m) sections ($n = 45$ cells). **f**, Schematic of a tibia showing the separation of epiphysis/metaphysis from diaphysis (red line). **g**, The frequency of CD150⁺CD48⁻LSK cells and α -catulin-GFP⁺c-kit⁺ cells by flow cytometry in the epiphysis/metaphysis versus diaphysis of tibias

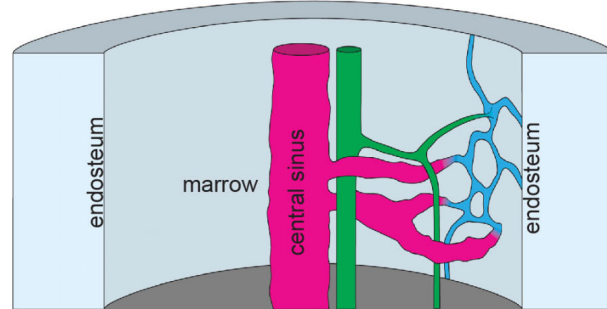
($n = 9$ mice in 2 independent experiments). **h**, The frequency of α -catulin-GFP⁺c-kit⁺ cells in the tibia epiphysis/metaphysis versus diaphysis based on deep confocal imaging ($n = 3$ bisected tibias in 3 independent experiments). **i**, Deep imaging of a bisected tibia showing the separation of metaphysis and diaphysis (red line) where the central sinus branches. Note that these tibias were digitally reconstructed from two different imaging sessions, above and below the diagonal white line. This image shows a 349μ m thick specimen collapsed into 2D. This causes α -catulin-GFP⁺ cells and c-kit⁺ cells to appear much more frequent than they actually were because all of the cells from the thick specimen were collapsed into a single 2D optical plane for presentation. **j**, For comparison purposes, a single 2μ m thick optical slice from the tibia in **i** is shown. **k**, High magnification images of single α -catulin-GFP⁺c-kit⁺ cells from the same tibia. Note that α -catulin-GFP is also expressed by sinusoidal endothelial cells but these cells are easily distinguished from HSCs because the endothelial cells lack c-kit expression and have a different morphology. Statistical significance was assessed using Student's *t*-tests (* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$). All data represent mean \pm s.d.



Extended Data Figure 8 | c-kit and α -catulin-GFP staining do not reflect autofluorescence or background staining; and GFAP⁺ non-myelinating Schwann cells tend to localize in the centre of the marrow. **a**, Four-colour confocal analysis of a bone marrow plug from a tibia diaphysis stained with primary and secondary antibodies against Ki-67, α -catulin-GFP, c-kit, and laminin. A 2 μ m optical section is shown from a thick specimen to illustrate typical staining. **b**, Negative control in which a bone marrow plug from a tibia diaphysis was stained with isotype control and secondary antibodies then imaged under the same conditions as shown in **a**. **c**, Ki-67 staining was largely or exclusively nuclear, co-localizing with DAPI. **d–g**, Low magnification images of bone marrow plugs from tibia diaphysis stained with antibodies against α -catulin-GFP, c-kit, and GFAP. GFAP⁺ non-myelinating Schwann cells are associated with nerve fibres that run longitudinally along the central bone marrow, where innervated arterioles are located²⁴. α -catulin-GFP⁺c-kit⁺ cells were identified and annotated with blue spheres using the Imaris spot function in **e** and **g**. Note, the blue spheres are larger than the actual HSCs because at their actual size, HSCs would be extremely difficult to see at this magnification. As the HSCs are represented as large blue spheres, they appear

denser than they actually are. For clarity, other haematopoietic cells and endothelial cells are not shown in **e** and **g**. **h**, A higher magnification image showing two α -catulin-GFP⁺c-kit⁺ cells (arrows) and their localization relative to GFAP positive glia (white) and α -catulin-GFP⁺ endothelial cells (green). The images in **d–g** show a 505 μ m thick specimen. This causes α -catulin-GFP⁺ cells and c-kit⁺ cells to appear more frequent than they actually were because all of the cells from the thick specimen were collapsed into a single 2D optical plane for presentation. Because these were thick specimens, there were cases in which an α -catulin-GFP⁺ cell and a c-kit⁺ cell were present in different optical planes such that they appeared to be a single α -catulin-GFP⁺c-kit⁺ cell when collapsed into a single 2D image. For this reason, α -catulin-GFP⁺c-kit⁺ cells cannot be reliably identified in low magnification 2D projected images. In all cases, cells that we identified as α -catulin-GFP⁺c-kit⁺ were manually examined at high magnification in 3D to confirm double labelling of single cells, as shown in **h**. Few HSCs were closely associated with nerve fibres in these images when they were examined at high magnification and in 3D.

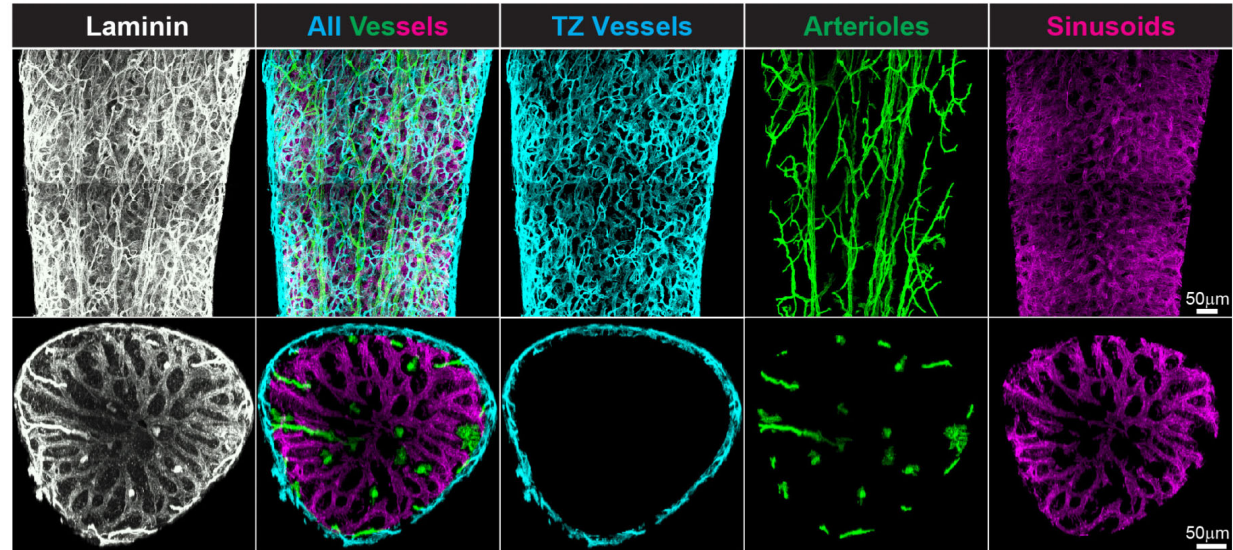
a Sinusoids Arterioles Transition Zone Vessels



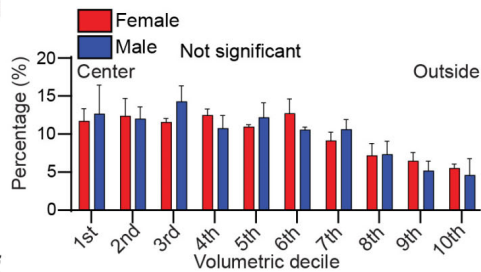
b

| | Vessel diameter | Basal lamina | Position |
|-------------------|-----------------|--------------|---|
| Arterioles | Variable | Continuous | Parallel to long axis |
| TZ Vessels | Small | Continuous | Outer 20% of marrow volume, close to bone surface |
| Sinusoids | Large | Fenestrated | Perpendicular to long axis |

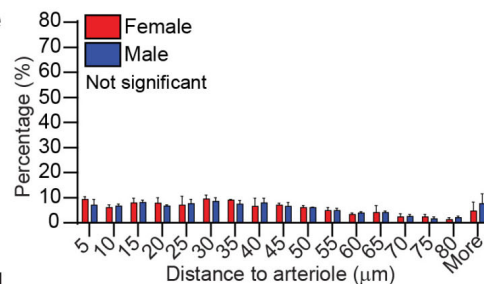
c



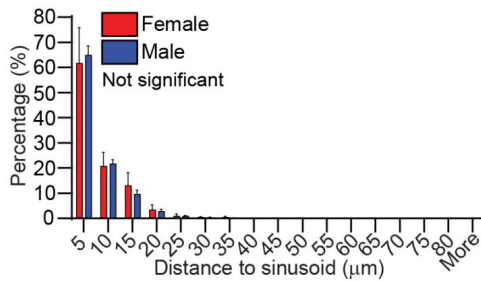
d



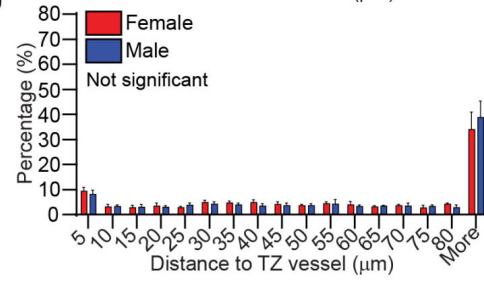
e



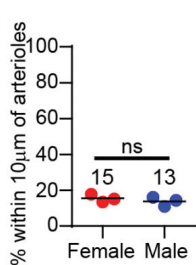
f



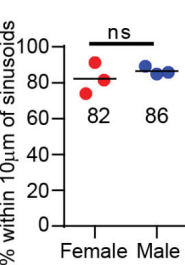
g



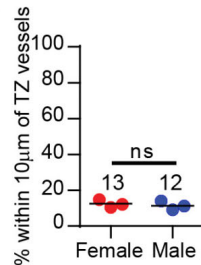
h



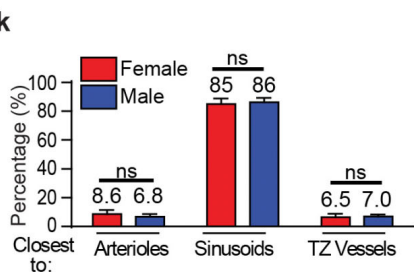
i



j

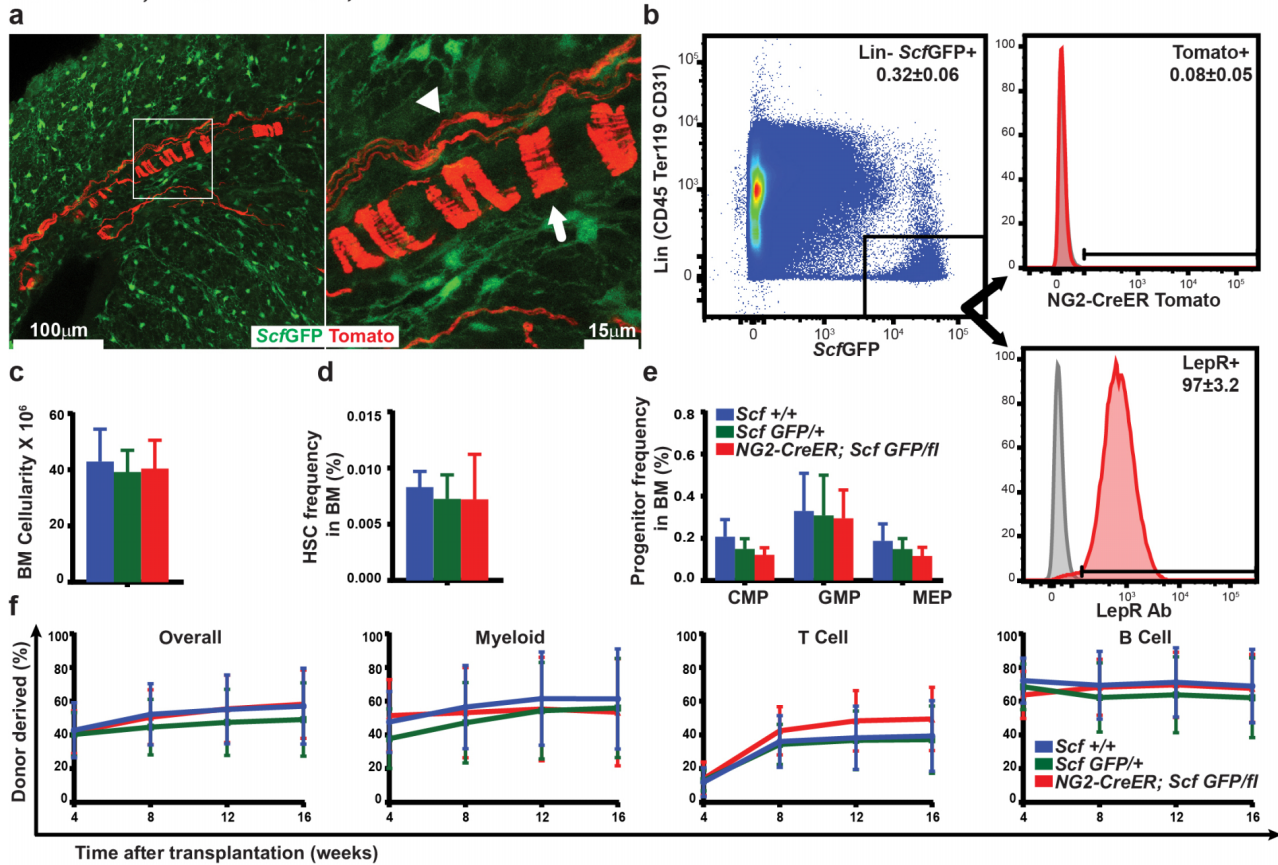
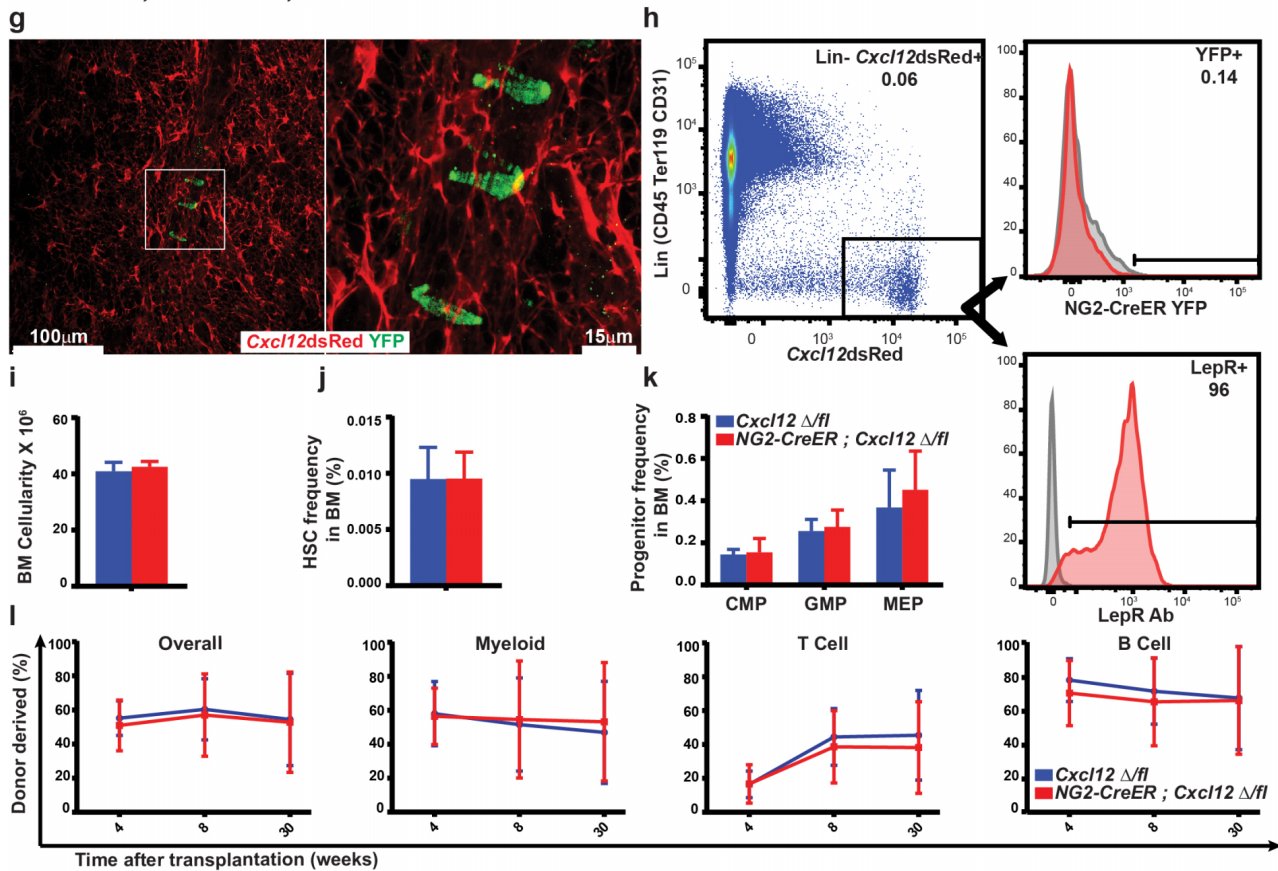


k



Extended Data Figure 9 | Bone marrow blood vessel types can be distinguished based on vessel diameter, continuity of basal lamina, morphology, and position; and no difference in the distribution of HSCs in the bone marrow of male and female mice was detected. **a, b**, Schematic (a) and properties (b) of blood vessels in the bone marrow. Blood enters the marrow through arterioles that branch as they become smaller in diameter and approach the endosteum, where they connect to smaller diameter transition zone capillaries near the bone surface. These transition zone capillaries connect to the large diameter sinusoids that feed blood into the central sinusoid through which it leaves the bone marrow in venous circulation. **c**, Each type of blood vessel was distinguished based on vessel diameter, continuity of basal lamina, morphology, and position, and then colour-coded using published criteria^{17,30,46,47}. To create distinct digital surfaces associated with each type of blood vessel, we first designated all laminin-stained blood vessels in the outer 20% of the marrow volume (adjacent to the endosteum) as transition zone (TZ) vessels (blue). Arterioles were identified and manually traced in the remaining 80% of marrow volume based on high intensity laminin staining, continuous basal lamina, and morphology. Remaining blood vessels with low

intensity laminin staining, fenestrated basal lamina, large diameter, and sinusoidal morphology were designated sinusoids. The longitudinal images (top) show bone marrow plugs that were 550 μm thick and the cross-sectional images (bottom) were 49 μm thick. **d**, The distribution of $\alpha\text{-catulin-GFP}^+\text{c-kit}^+$ cells in concentric cylinders corresponding to equal volumetric deciles from central marrow to endosteal marrow (near the bone surface) in bone marrow plugs from the tibia diaphysis of male and female mice. **e–g**, The distance from $\alpha\text{-catulin-GFP}^+\text{c-kit}^+$ cells in male or female mice to the nearest arteriole (e), sinusoid (f), or transition zone vessel (g) in tibia based on deep imaging. **h–j**, The percentage of $\alpha\text{-catulin-GFP}^+\text{c-kit}^+$ cells within 10 μm of arterioles (h), sinusoids (i) and transition zone vessels (j) in the tibias of male versus female mice. **k**, The percentage of $\alpha\text{-catulin-GFP}^+\text{c-kit}^+$ cells closest to arterioles, sinusoids, or transition zone vessels in the tibias of male versus female mice. These data show mean \pm s.d. for a total of 1,345 $\alpha\text{-catulin-GFP}^+\text{c-kit}^+$ cells from 3 female tibias and 1,632 $\alpha\text{-catulin-GFP}^+\text{c-kit}^+$ cells from 3 male tibias. The statistical significance of differences was assessed using Kolmogorov–Smirnov tests in **d–g** and Student's *t*-tests in **h–k**; none of the differences were statistically significant.

NG2-CreER ; Rosa tdTomato/+ ; Scf GFP/+*NG2-CreER ; Rosa YFP/+ ; Cxcl12 dsRed/+*

Extended Data Figure 10 | Expression of NG2-CreER was not detected in *Scf* or *Cxcl12*-expressing cells; and conditional deletion of *Scf* or *Cxcl12* using NG2-CreER did not affect HSC frequency or haematopoiesis. **a**, A 20 μm optical section from a 390- μm -thick cleared bone marrow plug from the tibia diaphysis of an *NG2-creER;Rosa^{tdTomato/+};Scf^{GFP/+}* mouse (image is representative of bones from 4 mice). The image shows rare tdTomato⁺ periarteriolar smooth muscle cells (arrow) as well as glia associated with nerve fibres (arrowhead); however, we were unable to detect *Scf* expression by any of these cells. **b**, Representative flow cytometry plots showing the percentage of *Scf*-GFP⁺ stromal cells that were positive for tdTomato expression (reflecting recombination by NG2-CreER) or Lepr antibody staining (mean \pm s.d. from 4 mice in 3 independent experiments). *Scf*-GFP⁺ stromal cells were uniformly positive for Lepr expression but negative for NG2-CreER recombination. **c–f**, Conditional deletion of *Scf* in *NG2-creER;Scf^{GFP/f}* mice had no effect on bone marrow cellularity (**c**), HSC frequency (**d**), CMP, GMP, or MEP frequency (**e**), or bone marrow reconstituting capacity upon transplantation

into irradiated mice (**f**) ($n = 5$ mice per genotype in 5 independent experiments with 4/5 recipient mice per donor in each experiment). **g**, A 20 μm optical section from the diaphysis of a 130- μm -thick cleared half tibia from an *NG2-creER;Rosa^{YFP/+};Cxcl12^{DsRed/+}* mouse. The image shows rare YFP⁺ periarteriolar smooth muscle cells; however, we were unable to detect *Cxcl12* expression by these cells. **h**, Representative flow cytometry plots showing the percentage of *Cxcl12*-DsRed⁺ stromal cells that were positive for YFP expression (reflecting recombination by NG2-CreER) or Lepr antibody staining. *Cxcl12*-DsRed⁺ stromal cells were uniformly positive for Lepr expression but negative for NG2-CreER. **i–l**, Conditional deletion of *Cxcl12* in *NG2-creER;Cxcl12^{-/f}* mice had no effect on bone marrow cellularity (**i**), HSC frequency (**j**), CMP, GMP, or MEP frequency (**k**), or bone marrow reconstituting capacity upon transplantation into irradiated mice (**l**) ($n = 4$ mice per genotype in 4 independent experiments with 4/5 recipient mice per donor in each experiment).

Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells

Devon A. Lawson^{1†}, Nirav R. Bhakta², Kai Kessenbrock^{1,3†}, Karin D. Prummel^{1†}, Ying Yu¹, Ken Takai^{1†}, Alicia Zhou³, Henok Eyob³, Sanjeev Balakrishnan³, Chih-Yang Wang^{1,4}, Paul Yaswen⁵, Andrei Goga^{2,3} & Zena Werb¹

Despite major advances in understanding the molecular and genetic basis of cancer, metastasis remains the cause of >90% of cancer-related mortality¹. Understanding metastasis initiation and progression is critical to developing new therapeutic strategies to treat and prevent metastatic disease. Prevailing theories hypothesize that metastases are seeded by rare tumour cells with unique properties, which may function like stem cells in their ability to initiate and propagate metastatic tumours^{2–5}. However, the identity of metastasis-initiating cells in human breast cancer remains elusive, and whether metastases are hierarchically organized is unknown². Here we show at the single-cell level that early stage metastatic cells possess a distinct stem-like gene expression signature. To identify and isolate metastatic cells from patient-derived xenograft models of human breast cancer, we developed a highly sensitive fluorescence-activated cell sorting (FACS)-based assay, which allowed us to enumerate metastatic cells in mouse peripheral tissues. We compared gene signatures in metastatic cells from tissues with low versus high metastatic burden. Metastatic cells from low-burden tissues were distinct owing to their increased expression of stem cell, epithelial-to-mesenchymal transition, pro-survival, and dormancy-associated genes. By contrast, metastatic cells from high-burden tissues were similar to primary tumour cells, which were more heterogeneous and expressed higher levels of luminal differentiation genes. Transplantation of stem-like metastatic cells from low-burden tissues showed that they have considerable tumour-initiating capacity, and can differentiate to produce luminal-like cancer cells. Progression to high metastatic burden was associated with increased proliferation and MYC expression, which could be attenuated by treatment with cyclin-dependent kinase (CDK) inhibitors. These findings support a hierarchical model for metastasis, in which metastases are initiated by stem-like cells that proliferate and differentiate to produce advanced metastatic disease.

To investigate differentiation in metastatic cells, we used a microfluidics-based platform (Fluidigm) for multiplex gene expression analysis in individual cells. This facilitated a systems-level approach to study the simultaneous expression of groups of genes and resolve cellular diversity during breast cancer metastasis only achievable at the single-cell level. We designed single-cell experiments to investigate 116 genes involved in stemness, pluripotency, epithelial-to-mesenchymal transition (EMT), mammary lineage specification, dormancy, cell cycle and proliferation (Supplementary Table 1)^{6–10}.

We first developed a single-cell gene expression signature from normal human breast epithelium to generate a reference for analysing differentiation in metastatic cells. The breast contains two epithelial lineages: the basal/myoepithelial lineage that contains stem cells, and a luminal lineage that contains progenitor and mature cell populations. We sorted single basal/stem, luminal, and luminal progenitor cells

from reduction mammoplasty samples from three individuals, and processed them according to established protocols (Fig. 1a)^{10–13}. Principal component analysis (PCA) and unsupervised hierarchical clustering showed that basal and luminal cells represent distinct populations in each individual, as expected (Fig. 1b, d). Forty-nine of the one-hundred and sixteen genes tested showed differential expression between basal/stem and luminal cells, and were used to generate a 49-gene differentiation signature. This signature included established lineage-specific genes such as *KRT5*, *TP63*, *MUC1*, *CD24* and *GATA3* (Fig. 1c, d, Supplementary Table 2 and Supplementary Data 1), validating our multiplex quantitative polymerase chain reaction (qPCR) approach.

Mice from three genetically distinct triple-negative (ER[−]PR[−]HER2[−]), basal-like patient-derived xenograft (PDX) models (HCI-001, HCI-002 and HCI-010) were analysed (Extended Data Table 1)¹⁴. We focused on this subtype since it is the most aggressive, metastasis is frequent, and there are no targeted therapeutics to treat it¹⁵. These PDX models maintain the essential properties of the original patient tumours, including metastatic tropism, making them authentic experimental systems for studying human cancer metastasis¹⁴.

To isolate metastatic cells from PDX mice, we first developed a highly sensitive, species-specific FACS-based assay. We annotated published microarray data to identify cell surface genes highly expressed in PDX breast cancer cells¹⁴. This revealed as a top candidate CD298 (also known as *ATP1B3*), which is a β -subunit of the Na⁺/K⁺ ATPases that are essential for basic cellular function¹⁶. Using a human species-specific antibody, we found that CD298 is expressed by >99.9% of cells in three different human mammary cell lines, with no background in mouse lines or control mouse peripheral tissues (Fig. 2b and Extended Data Fig. 1a, b). In dissociated PDX primary tumours, all cells either expressed human CD298 or mouse major histocompatibility complex class I (MHC I), indicating that CD298 could detect nearly all cells (>99.5%) that were not of mouse origin (Fig. 2a). We therefore expected that this assay would capture the majority of metastatic cells in PDX mice, with negligible false-positive rates. CD298 was also superior to commonly used markers, such as human EpCAM, CD24 and MHC I (Extended Data Fig. 1c).

We detected metastatic cells in peripheral tissues of 70/100 (70%) PDX mice using this assay, including the lung, lymph node, bone marrow, liver, brain and peripheral blood (Extended Data Table 1). All animals were analysed when their primary tumour reached 20–25 mm in diameter, and primary tumour growth kinetics were consistent within each model (Extended Data Fig. 2a–d). Although animals were analysed at the same endpoint, we observed variation in metastatic burden by FACS and histology (Fig. 2b, c). We exploited this to investigate gene expression in advanced-stage metastatic disease (high burden) versus earlier-stage metastatic disease (low burden). In total we analysed over 20 mice, and show comprehensive analysis of 441 metastatic and 523 primary tumour cells from 12 animals. The tissues were rank ordered by burden, from

¹Department of Anatomy, University of California, San Francisco, California 94143, USA. ²Department of Medicine, University of California, San Francisco, California 94143, USA. ³Department of Cell and Tissue Biology, University of California, San Francisco, California 94143, USA. ⁴Institute of Basic Medical Sciences, College of Medicine, National Cheng Kung University, Tainan 70101, Taiwan. ⁵Department of Cell and Molecular Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [†]Present addresses: Department of Physiology and Biophysics, University of California, Irvine, California 92697, USA (D.A.L.); Department of Biological Chemistry, University of California, Irvine, California 92697, USA (K.K.); Institute of Molecular Life Sciences, University of Zürich, Zürich 8057, Switzerland (K.D.P.); Saitama Cancer Center, Saitama 362-0806, Japan (K.T.).

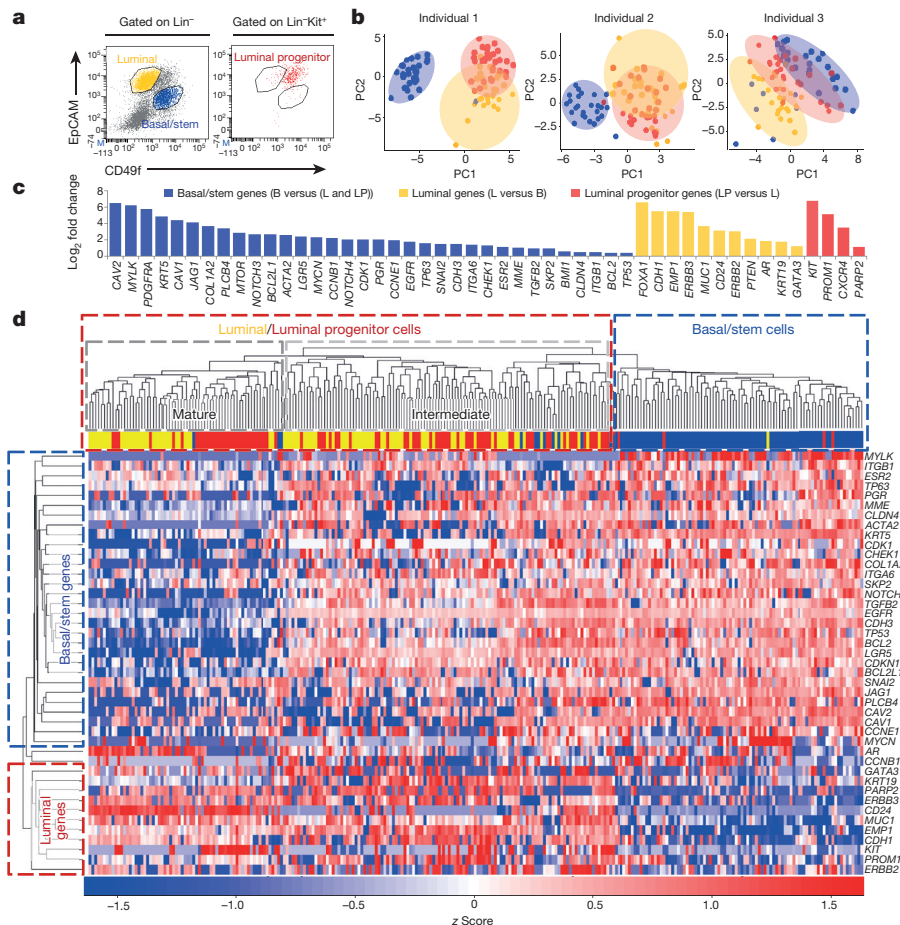


Figure 1 | Single-cell analysis of normal human mammary epithelial cells. **a**, FACS plots show basal/stem ($\text{Lin}^- \text{CD49f}^{\text{hi}} \text{EpCAM}^{\text{lo}} \text{cKit}^-$, blue), luminal ($\text{Lin}^- \text{CD49f}^{\text{lo}} \text{EpCAM}^{\text{hi}} \text{cKit}^-$, yellow), and luminal progenitor ($\text{Lin}^- \text{CD49f}^{\text{med}} \text{EpCAM}^{\text{med}} \text{cKit}^+$, red) cells from a representative mammaplasty patient. $\text{Lin} = \text{CD45/CD31}$. **b**, PCA plots show distinct cell populations identified in three patients. PC, principal component. **c**, Bar graph shows the 49 of 116 genes that were significantly ($P < 0.05$) differentially expressed between the populations. P values and fold change are listed in Supplementary Table 2. B, basal/stem; LP, luminal progenitor; L, luminal. **d**, Heatmap and dendrogram show unsupervised hierarchical clustering of individual cells and genes from the 49-gene signature that were run on all arrays.

lowest (light grey) to highest (black) (Extended Data Fig. 2e). Circulating tumour cells (CTCs) in the blood, and disseminated tumour cells (DTCs) in the bone marrow were not included in the ranking since overt metastasis was never observed in these tissues.

Remarkably, PCA plots for individual animals showed that in tissues with low burden, metastatic cells were very distinct from the primary tumour cells they were derived from (Fig. 3a). By contrast, metastatic cells from higher burden animals were more similar to primary tumour cells. This was also observed by unsupervised hierarchical clustering of pooled cells from all animals, which showed that low-burden metastatic cells form a unique cluster, while higher-burden metastatic cells cluster with primary tumour cells (Extended Data Fig. 3a). Most strikingly, we found that this was due to a conserved basal/stem-cell signature in low-burden metastatic cells across all animals and models. Analysis of genes comprising the 49-gene differentiation signature showed that low-burden metastatic cells expressed higher levels of 22 basal/stem-cell genes, including *LGR5*, *BMII*, *BCL2*, *NOTCH4* and *JAG1*, and lower levels of seven luminal genes, such as *MUC1*, *EMP1* and *CD24* (Fig. 3b). Focusing on clustering of only the metastatic cells (Fig. 3c), we discovered considerable heterogeneity in differentiation, which directly correlated with metastatic burden. Akin to the normal mammary gland, metastatic cells organized into two distinct clusters, where low-burden metastatic cells were most basal/stem-like, and higher-burden metastatic cells possessed a spectrum of progressively more luminal-like expression patterns. This was also observed when lung metastatic cells from each PDX model were analysed separately (Extended Data Fig. 4a and Supplementary Data 2), indicating that it is a conserved phenomenon in each model. Some differences in gene expression were observed between lung metastatic cells from different patient models, but they were not sufficient to cluster cells separately by PDX model (Extended Data Fig. 4c, d and Supplementary Data 3).

To investigate heterogeneity at the protein level, we performed immunostaining for KRT5 (basal) and MUC1 (luminal) (Extended Data Fig. 4e). Tumour cells found in micrometastases from low-burden tissues were largely KRT5⁺ (95.8%) and MUC1⁻ (94.3%), while cells from high-burden tissues were heterogeneous for KRT5 and largely MUC1⁺ (72.9%). This suggests that differentiation status also correlates with metastatic burden at the protein level.

By single-cell analysis, low-burden metastatic cells expressed very high levels of the pluripotency genes *POU5F1* (also known as *OCT4*) and *SOX2*, suggesting that they may exploit embryonic programs for self-renewal and maintenance (Fig. 3b). Low-burden metastatic cells also expressed higher levels of typical EMT markers such as *SNAIL2*, *SKP2* and *TWIST1*, and lower levels of *CDH1*, which was observed in normal basal/stem cells (with the exception of *TWIST1*) (Fig. 3b and Extended Data Table 2). This is consistent with previous reports showing that EMT promotes stemness in the mammary gland, and suggests that low-burden metastatic cells may utilize an EMT program to facilitate dissemination^{17,18}. Gene ontology enrichment revealed that genes involved in the DNA damage response, chromatin modification, differentiation, apoptosis and the cell cycle were differentially expressed in low-burden metastatic cells (Supplementary Data 4). Extended Data Table 2 and Supplementary Data 5 list all 55 genes (of 116 analysed) that were differentially expressed in low-burden metastatic cells.

The heterogeneity observed in metastatic cells raised the question of whether stem-like metastatic cells directly give rise to luminal-like cells, or whether they originate from distinct founder cells. To test first whether cells that disseminate at early phases of primary tumour growth can produce luminal-like metastatic cells, we resected primary tumours when they were only 10–12 mm in diameter and allowed metastases to grow for 8 weeks. Single-cell analysis of the resulting lung metastatic cells showed that 85.4% were luminal-like, and clustered with high-burden metastatic cells from previous experiments

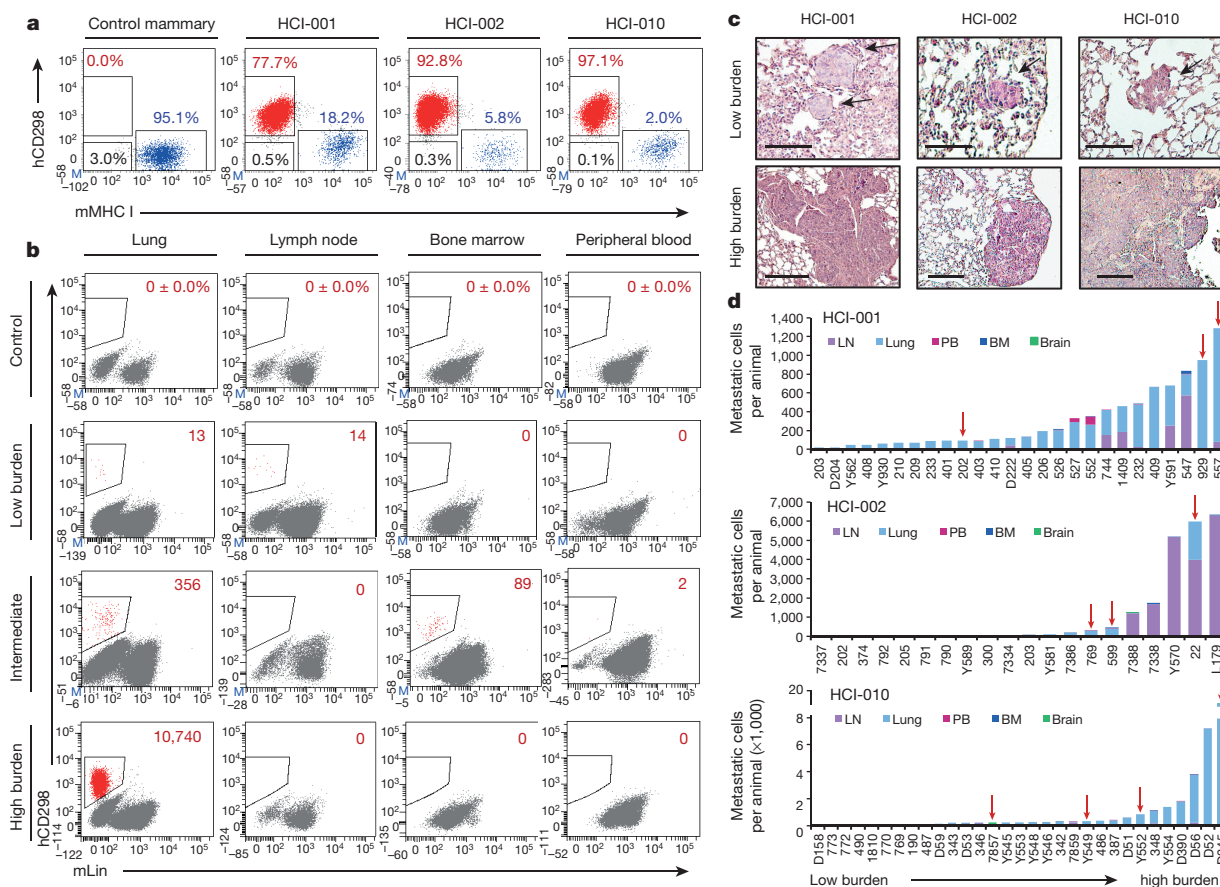


Figure 2 | Identification of human metastatic cells in PDX mice. **a**, FACS plots show human (h)CD298⁺ (red), mouse (m)MHC I⁺ (blue), and double-negative (black) cells in representative tissues ($n = 3$). **b**, FACS plots show percentage or number of hCD298⁺mLn[−] (mTer119/mCD45/mCD31) cells in representative low- and high-burden mice. **c**, Haematoxylin and eosin stains show micro- and macrometastatic lesions in lung tissues of low- and high-

burden mice. Low-burden scale bar, 100 μ m; high-burden scale bar, 200 μ m. Arrows indicate micrometastatic lesions. **d**, Histograms show the distribution of metastatic burden in each model. Only animals with metastases are shown. Red arrows indicate animals subjected to single-cell analysis. BM, bone marrow; LN, lymph node; PB, peripheral blood.

(Extended Data Fig. 4b). This suggests that luminal-like metastases can derive from cells that disseminate at earlier stages of primary tumour growth.

To test the growth and differentiation capacity of stem-like metastatic cells directly, we transplanted low-burden metastatic cells into mammary glands. Remarkably, two of four transplants produced large tumours (Extended Data Fig. 5a), by contrast with primary tumour cells, which did not produce tumours even at 100-fold higher numbers. This is consistent with previous reports indicating that PDX tumours are more efficiently propagated as fragments than dissociated cells¹⁹. Single-cell analysis of the resulting tumour cells showed that 98.7% of them were luminal-like, and clustered with primary tumour cells and high-burden metastatic cells from previous experiments (Extended Data Fig. 5b). This suggests that low-burden metastatic cells have considerable tumour-initiating capacity, and can give rise to luminal-like tumour cells, supporting the hypothesis that stem-like metastatic cells give rise to luminal-like ones.

A compelling question raised in this study is whether stem-like cells are present in primary tumours, or whether they evolve later through interaction with their new microenvironment. Unsupervised hierarchical clustering shows that 1.4% of primary tumour cells cluster with low-burden metastatic cells and possess a basal/stem-like phenotype (Extended Data Fig. 3a). This is consistent with previous findings that rare invasive 'leader' cells on the periphery of primary tumours express basal cell markers²⁰. Interestingly, the most metastatic PDX model (HCI-010) had the highest percentage of basal/stem-like primary tumour cells, while the least metastatic model (HCI-002) had the low-

est. This suggests that primary tumours contain a rare subpopulation of stem-like cells, and that the percentage correlates with metastatic potential. This led us to investigate whether enrichment of this stem-like signature in primary tumours may be predictive of distant metastasis in human patient data sets. By Kaplan–Meier analysis, we found that 16 of 55 genes associated with stem-like metastatic cells were significantly prognostic (Supplementary Data 6). Future studies to determine whether the frequency of stem-like cells in primary tumours can be used as a predictive biomarker for metastasis may be clinically valuable.

Previous work has shown that metastatic cells in different organs display distinct gene expression signatures². Consistent with this, by supervised clustering of cells by target organ, we found that metastatic cells in the brain, bone marrow and peripheral blood displayed distinct gene expression patterns (Extended Data Fig. 6a). Brain metastatic cells were the most distinct, and expressed the highest levels of stem cell, quiescence and anti-apoptosis genes. In total, 80 genes were significantly differentially expressed between the populations (Extended Data Fig. 6b, Supplementary Table 3 and Supplementary Data 7).

CTCs are of particular clinical interest for use as a 'liquid biopsy' for diagnosis and prognosis. Although only rare CTCs could be recovered, they most closely resembled lung metastatic cells, and were least similar to brain metastatic cells (Extended Data Fig. 6c). Interestingly, most CTCs and bone marrow DTCs clustered with 'intermediate' metastatic cells, which may be because the cells were harvested from animals with intermediate burden (Extended Data Fig. 2e). However, 16.7% and 10.7%, respectively, showed a more basal/stem-like

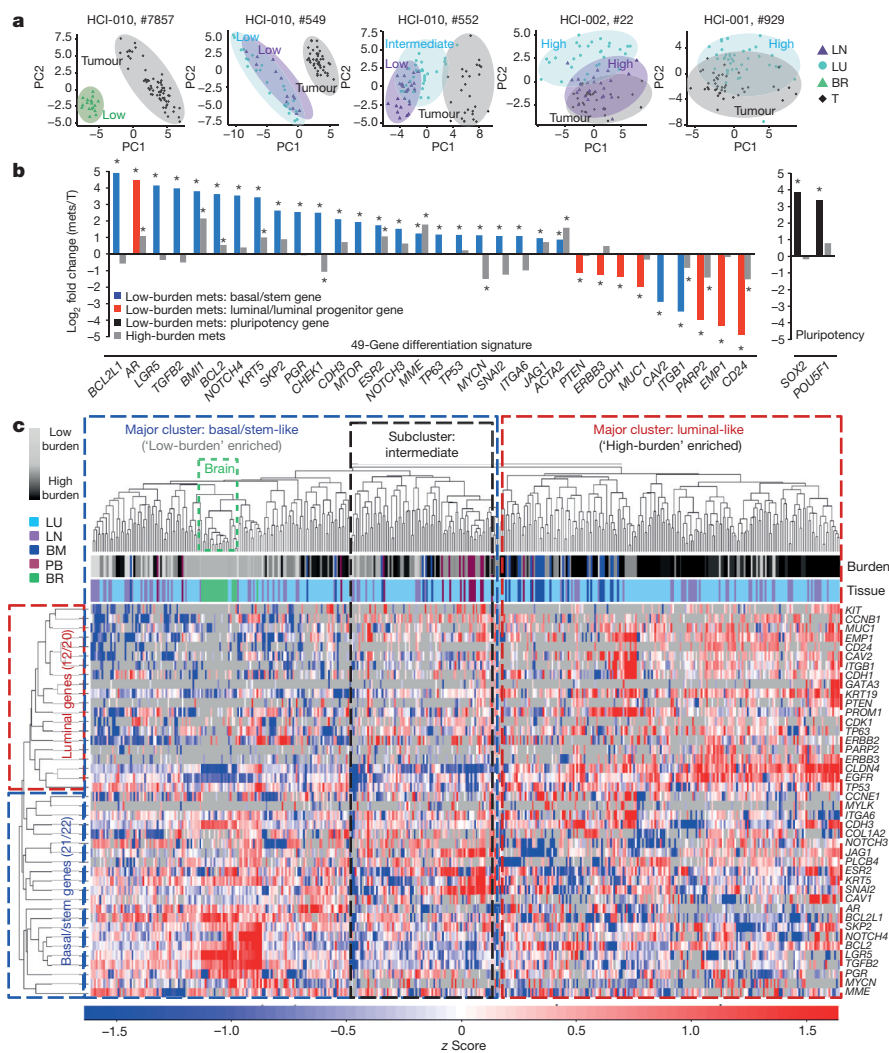


Figure 3 | Early stage metastatic cells possess a distinct basal/stem-cell program. **a**, PCA plots show metastatic and primary tumour cells in representative mice. Low, high and intermediate indicate burden levels. **b**, Bar graph shows genes from the 49-gene differentiation signature, and pluripotency genes, that were differentially expressed in low-burden metastatic cells. $*P < 0.05$, significant relative to primary tumour; primary tumour expression = 0. P values and fold change are listed in Extended Data Table 2. Mets, metastases. **c**, Heatmap and dendrogram show unsupervised hierarchical clustering of metastatic cells and genes from the 49-gene signature that were run on all arrays. BM, bone marrow; BR, brain; LN, lymph node; LU, lung; PB, peripheral blood; T, tumour.

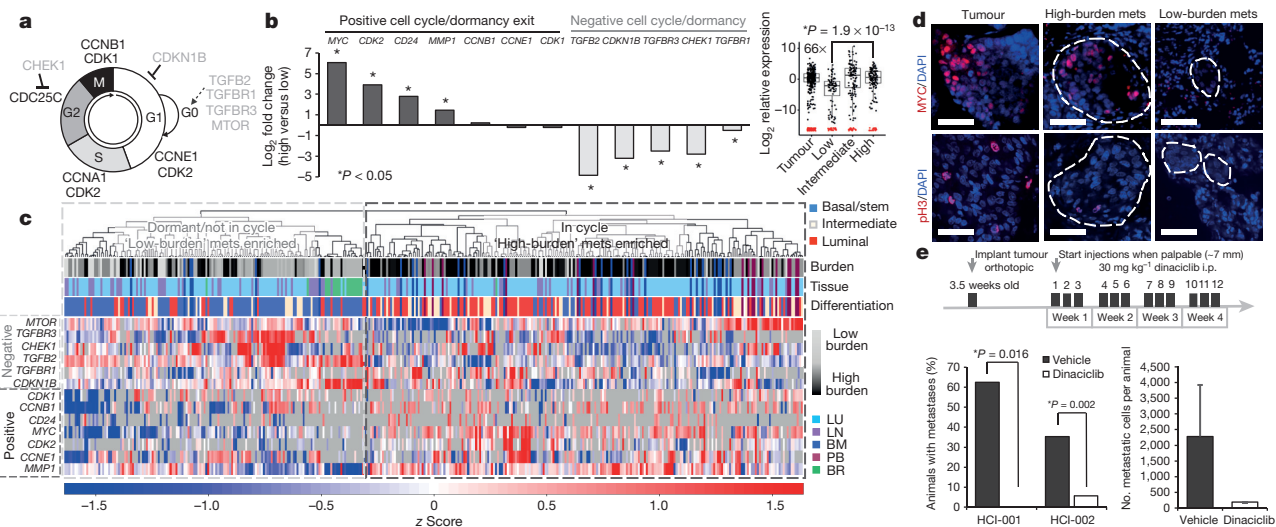


Figure 4 | Metastatic progression is blocked by cell cycle inhibition. **a**, Schematic of the cell cycle. **b**, Graph and box plot (*MYC*) show expression for dormancy and cell-cycle-associated genes. P values: *MYC*: 1.9×10^{-13} ; *CDK2*: 1.25×10^{-7} ; *CD24*: 0.005; *MMP1*: 1.08×10^{-12} ; *TGFB2*: 1.28×10^{-16} ; *CDKN1B*: 2.63×10^{-14} ; *CHEK1*: 2.18×10^{-6} . **c**, Unsupervised hierarchical clustering of metastatic cells and cell-cycle-associated genes. mets, metastases.

d, Immunofluorescence stains for MYC and phospho-histone H3 (pH3) in micro- and macrometastatic lesions. Scale bar, 50 μ m. **e**, Dinaciclib treatment course in PDX mice (top). Graphs show percentage of mice with metastasis, and burden per animal. Error bars, s.d. Only one drug-treated animal developed metastasis, so no P value was generated (right graph).

signature (Fig. 3c, basal/stem-like cluster), suggesting that these stem-like cells may represent the true metastatic seeder cells.

We also observed a shift towards a more proliferative signature associated with increased metastatic burden. Low-burden metastatic cells expressed higher levels of quiescence and dormancy-associated genes, including *CDKN1B*, *CHEK1*, *TGFBR3* and *TGFB2* (Fig. 4a, b)^{21,22}. Higher-burden metastatic cells appeared to enter the cell cycle, expressing lower levels of quiescence and dormancy-associated genes and higher levels of cell-cycle-promoting genes such as *MYC* and *CDK2*, as well as *MMP1* and *CD24*, which have been associated with reactivation after dormancy. This distinction was further corroborated by unsupervised hierarchical clustering, showing that low- and high-burden metastatic cells form distinct clusters based on differential expression of these genes (Fig. 4c). Of note, the majority of metastatic cells in the dormant cluster were also in the basal/stem-cell cluster depicted in Fig. 3c, demonstrating a correlation between dormancy and stem-cell-related gene expression in metastatic cells. We also detected primary tumour cells (22.2%) with this less-proliferative signature (Extended Data Fig. 3b). Immunostaining for *MYC*, phospho-histone H3 and Ki67 confirmed that micrometastases show lower *MYC* expression and proliferative index (Fig. 4d and Extended Data Fig. 7a, b).

These findings prompted us to test whether blocking this switch from dormancy into the cell cycle could inhibit metastatic progression. Since we observed high levels of both *MYC* and *CDK2* in more advanced stage metastatic cells (Fig. 4b), we chose to test dinaciclib, a CDK inhibitor that has been shown to induce apoptosis in high *MYC*-expressing cancer cells via synthetic lethality^{23,24}. We hypothesized that apoptosis would be induced in metastatic cells transitioning into proliferation, since they appear to upregulate *MYC*. We administered dinaciclib to a total of 49 mice from two PDX models, HCI-001 and HCI-002, which were from drug-naïve patients. After a 4-week treatment course, we found that only 1 of 24 drug-treated animals displayed metastatic cells, in comparison to 44% (11/25) of vehicle-treated mice (Fig. 4e). Although tumour growth was delayed in drug-treated animals, many developed sizeable tumours by the endpoint, suggesting that the effect was not simply due to inhibition of the primary tumour (Extended Data Fig. 7c–e). By looking in high resolution at gene expression in single metastatic cells, we have uncovered previously unrealized diversity in differentiation and gene expression relating to the metastatic stage (Extended Data Fig. 8), and demonstrate that this approach can facilitate the identification of new potential drug targets with efficacy against metastatic disease.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 August 2014; accepted 29 July 2015.

Published online 23 September 2015.

1. Weigelt, B., Peterse, J. L. & van't Veer, L. J. Breast cancer metastasis: markers and models. *Nature Rev. Cancer* **5**, 591–602 (2005).
2. Oskarsson, T., Batlle, E. & Massagué, J. Metastatic stem cells: sources, niches, and vital pathways. *Cell Stem Cell* **14**, 306–321 (2014).
3. Hermann, P. C. *et al.* Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. *Cell Stem Cell* **1**, 313–323 (2007).
4. Pang, R. *et al.* A subpopulation of CD26⁺ cancer stem cells with metastatic capacity in human colorectal cancer. *Cell Stem Cell* **6**, 603–615 (2010).
5. Dieter, S. M. *et al.* Distinct types of tumor-initiating cells form human colon cancer tumors and metastases. *Cell Stem Cell* **9**, 357–365 (2011).
6. Grigoriadis, A. *et al.* Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data. *Breast Cancer Res.* **8**, R56 (2006).

7. Jones, C. *et al.* Expression profiling of purified normal human luminal and myoepithelial breast cells: identification of novel prognostic markers for breast cancer. *Cancer Res.* **64**, 3037–3045 (2004).
8. Kendrick, H. *et al.* Transcriptome analysis of mammary epithelial subpopulations identifies novel determinants of lineage commitment and cell fate. *BMC Genomics* **9**, 591 (2008).
9. Raouf, A. *et al.* Transcriptome analysis of the normal human mammary cell commitment and differentiation process. *Cell Stem Cell* **3**, 109–118 (2008).
10. Shehata, M. *et al.* Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res.* **14**, R134 (2012).
11. Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84–88 (2006).
12. Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993–997 (2006).
13. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nature Med.* **15**, 907–913 (2009).
14. DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nature Med.* **17**, 1514–1520 (2011).
15. Dent, R. *et al.* Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin. Cancer Res.* **13**, 4429–4434 (2007).
16. Malik, N., Canfield, V. A., Beckers, M. C., Gros, P. & Levenson, R. Identification of the mammalian Na,K-ATPase 3 subunit. *J. Biol. Chem.* **271**, 22754–22758 (1996).
17. Mani, S. A. *et al.* The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **133**, 704–715 (2008).
18. Guo, W. *et al.* Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* **148**, 1015–1028 (2012).
19. Landis, M. D., Lehmann, B. D., Pietersen, J. A. & Chang, J. C. Patient-derived breast tumor xenografts facilitating personalized cancer therapy. *Breast Cancer Res.* **15**, 201 (2013).
20. Cheung, K. J., Gabrielson, E., Werb, Z. & Ewald, A. J. Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell* **155**, 1639–1651 (2013).
21. Bragado, P. *et al.* TGF- β 2 dictates disseminated tumour cell fate in target organs through TGF- β -RIII and p38 α / β signalling. *Nature Cell Biol.* **15**, 1351–1361 (2013).
22. Kim, R. S. *et al.* Dormancy signatures and metastasis in estrogen receptor positive and negative breast cancer. *PLoS ONE* **7**, e35569 (2012).
23. Horiuchi, D. *et al.* MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition. *J. Exp. Med.* **209**, 679–696 (2012).
24. Huskey, N. E. *et al.* CDK1 inhibition targets the p53-NOXA-MCL1 axis, selectively kills embryonic stem cells, and prevents teratoma formation. *Stem Cell Reports* **4**, 374–389 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Welm for providing access to PDX tissues developed by her group, which served as the foundation for this study. We also thank K. Lee, R. Kumar, A. Le, R. Daneman, J. Stingl and M. Binneweis for comments and technical contributions. This study was supported by funds from the National Cancer Institute (CA180039 and CA136717), Stand Up To Cancer/AACR (DT0409), the Era of Hope Scholar Award (W81XWH-12-1-0272), the Breast Cancer Research Foundation and the Atwater Foundation, and D. and J. Vander Wall. D.A.L. was supported by a US Department of Defense Congressionally Directed Medical Research Program postdoctoral fellowship (11-1-0742), and C.W. is supported by a grant from the Ministry of Science and Technology, Taiwan (104-2917-I-006-002).

Author Contributions K.T. initiated the PDX models, and along with D.A.L., Y.Y., H.E. and A.Z. performed transplants and maintained serial passages of PDX models. D.A.L., K.D.P., and Y.Y. harvested and analysed PDX tissues. K.D.P. performed histological analysis of PDX mouse tissues. D.A.L., K.D.P., Y.Y., A.Z. and H.E. performed dinaciclib treatment experiments. K.K. performed dinaciclib experiments. P.Y. prepared reduction mammaplasty samples. D.A.L. isolated cells by FACS and performed single-cell dynamic array experiments. N.R.B. designed algorithms for single-cell qPCR analyses in R and contributed to multiplex PCR experimental design. D.A.L. and N.R.B. performed analyses in R. D.A.L. wrote the manuscript, and with K.K. designed figures and schematics. C.-Y.W. and S.B. performed bioinformatics analyses. All authors contributed to experimental design and conceived experiments. A.G. and Z.W. provided overall guidance, funding and assisted in manuscript completion.

Author Information Single-cell multiplex qPCR data have been deposited in the Gene Expression Omnibus under accession number GSE70555. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.W. (zena.werb@ucsf.edu) or A.G. (andrei.goga@ucsf.edu).

METHODS

Cell line and animal experiments. All cell lines used in the study were pre-validated and grown using standard protocols that can be found on the American Type Culture Collection. The University of California, San Francisco Institutional Animal Care and Use Committee (IACUC) reviewed and approved all animal experiments. PDX tumour tissues were acquired from the laboratory of A. Welm and serially passaged as $\sim 8 \text{ mm}^3$ tumour fragments into the cleared inguinal fat pads of pre-pubescent NOD/SCID mice following established protocols¹⁴. When tumours became palpable, they were calipered weekly to monitor growth kinetics. Tumour fragments were stored by freezing in 90% FBS and 10% dimethylsulfoxide (DMSO) in liquid nitrogen. Clinical details of patients used for generation of each PDX model are detailed elsewhere¹⁴. All PDX animals were euthanized at the endpoint unless otherwise noted, when tumours reached 20–25 mm. In resection experiments, tumours were surgically removed at 10–12 mm. Resected animals were replaced in the colony and allowed to grow metastases for 8 weeks, at which time lung tissues were harvested, digested, and analysed by FACS for human cells.

For orthotopic transplant experiments for functional activity of metastatic cells, lymph node metastatic cells from animals with $<500 \text{ CD298}^+$ metastatic cells in the lymph nodes were isolated by FACS and pooled from several animals. CD298^+ primary tumour cells from matched animals were also isolated by FACS. Sorted cells were pelleted and resuspended in 1:1 Matrigel plus DMEM/F12 media. Sample dilutions were injected into cleared mammary fat pads of 3.5-week-old NOD/SCID mice. Grafts were harvested 4.5 months later when primary tumours reached 20 mm.

Dinaciclib treatment experiments. Dinaciclib was prepared and administered according to previously established protocols in mice^{23,25}. Dinaciclib was reconstituted in 20% hydroxypropyl β cyclodextrin (HPBCD). Animals were randomly assigned into treatment or control groups when tumour cells were transplanted, and mice were analysed using a single-blind design. The drug treatment course was initiated when tumours became palpable. A total of 49 animals (HCI-001 and HCI-010) were treated by i.p. injection three times per week at 30 mg kg^{-1} of drug, or vehicle (HPBCD), a previously established dose in mice²⁵. Animal group size was chosen by power analysis, using a two-tailed α of 0.05 with 80% power, and the frequencies of metastasis that we observed in each model (Extended Data Table 1). Animals were measured by caliper twice weekly to record primary tumour growth. Mice were euthanized at the conclusion of a 4-week treatment course, or earlier if their tumours reached the IACUC-established ethical endpoint (20 mm in diameter). Animals that developed adverse effects (for example, $>20\%$ weight loss) were excluded from the study. Statistical significance between drug- and vehicle-treated groups was examined by two-tailed, unpaired *t*-tests.

Bioinformatics and computational analysis of microarray data sets. Published microarray data sets (Gene Expression Omnibus (GEO) accession number GSE32531) on the PDX models were downloaded from the GEO database¹⁴. Microarray gene expression values were calculated by global median normalization and annotated with GeneSpring GX 12.0 software (Agilent Technologies). Plasma membrane genes highly expressed across all 15 PDX tumour samples and 12 original patient tumour samples included in the study were ranked ordered from highest to lowest expression across all the samples using the GENE-E package²⁶.

The prognostic value of each of the 55 genes characteristic of low-burden metastatic cells (Extended Data Table 2) was determined by Kaplan–Meier analysis using KM-plotter online software (<http://kmplot.com/analysis/>)²⁷. The relationship of gene expression and distant metastasis-free survival (DMFS) ($n = 1,610$) was evaluated in an integrated multi-study breast cancer microarray data set containing 13 breast cancer expression profiling data sets from GEO. Kaplan–Meier estimates of DMFS were calculated by setting the software to look for the optimal cut-off for separation of patients into high- and low-expressing groups. The hazard ratio, log-rank *P* value, and number of patients in each group are shown on the KM plot for each gene.

Tissue dissociation. All solid tissues, including primary tumour, liver, lungs, lymph nodes (axillary, brachial, cervical, sciatic and lumbar) and brain were dissociated for FACS using the same protocol. Briefly, tissues were mechanically chopped with scalpels, placed in culture medium (DMEM/F12 with 5% FBS, $5 \mu\text{g ml}^{-1}$ insulin (UCSF Cell Culture Facility), 50 ng ml^{-1} gentamycin (UCSF Cell Culture Facility) containing 2 mg ml^{-1} collagenase-1 (Sigma). They were then digested for 45 min at 37°C . The resulting suspensions were resuspended in $2 \text{ U } \mu\text{l}^{-1}$ DNase for 3 min at room temperature, washed and dissociated with 2 ml of 0.05% trypsin/EDTA (UCSF Cell Culture Facility) for 10 min at 37°C . Peripheral blood was collected by effusion with 10 mM EDTA in D-PBS, followed by mixture with 2% dextran in D-PBS for sedimentation of red blood cells using standard methods. After 1 h, supernatant was collected and cells were pelleted at $1,500 \text{ r.p.m.}$ for 5 min. Bone marrow was collected by removing all tissue from femur and tibia and flushing marrow with $1\times$ PBS using a 27G needle. Residual erythrocytes in peripheral blood, lung and tumour samples

were lysed with Red Blood Cell Lysis Buffer for 5 min at room temperature. All samples not used immediately were filtered through a $70 \mu\text{m}$ filter, and frozen in DMEM/F12 with 50% serum and 10% DMSO, and stored in liquid N_2 .

Reduction mammaplasty samples were acquired from the Cooperative Human Tissue Network (CHTN), a program funded by the National Cancer Institute. Tissues were washed three to five times with PBSA ($1\times$ Dulbecco's PBS supplemented with 200 U ml^{-1} penicillin, $200 \mu\text{g ml}^{-1}$ streptomycin (Invitrogen) and $5 \mu\text{g ml}^{-1}$ Fungizone (Invitrogen)). Tissues were minced into small fragments and digested overnight in collagenase-I-containing solution as previously described²⁸. Digested organoids were pelleted in a centrifuge at $100g$ for 3 min and frozen and stored in liquid N_2 as described earlier.

Flow cytometry. Antibodies for the human antigens CD45 (Alexa-450, eBioscience), CD31 (Alexa-450, eBioscience), CD298 (PE, Biolegend), EpCAM (PE or APC, eBioscience), CD49f (APC, eBioscience), CD117/cKit (FITC, eBioscience), CD24 (APC, eBioscience) and MHC I (APC, eBioscience) were purchased commercially. For mouse antigens, CD45 (FITC, eBioscience), Ter119 (FITC, eBioscience), CD31 (FITC, eBioscience) and MHC I (APC, eBioscience) were used. All antibodies were validated in previous publications^{10–13}, or in this study directly (CD298). Antibody staining was performed in DMEM/5% FBS supplemented with penicillin and streptomycin. After 15 min on ice, stained cells were washed of excess unbound antibodies and resuspended in medium. Flow sorting was done using a BD FACSAriaII cell sorter (Becton Dickinson), and analysis was done on an LSRII (Becton Dickinson). Forward-scatter height versus forward-scatter width (FSC-H versus FSC-W) and side-scatter area versus side-scatter width (SSC-A versus SSC-W) were used to eliminate cell aggregates and ensure single cell sorting. Dead cells were eliminated by excluding Sytox positive (SYTOX Blue dead cell stain, Molecular Probes) cells, which increased the efficiency of sorting robust, live cells for single-cell experiments. Contaminating human or mouse haematopoietic and endothelial cells were excluded by gating out Lin^+ (CD45, Ter119, CD31) cells. In Fig. 2a, $\text{Sytox}^+\text{mLin}^+$ cells were pre-gated out, and the percentages shown reflect the remaining population. Control mammary: $0.0 \pm 0.0\%$ hCD298^+ ; $95.1 \pm 2.0\%$ mMHC I^+ ; $3.0 \pm 2.1\%$ $\text{hCD298}^-\text{mMHC I}^+$; HCI-001: $77.7 \pm 11.3\%$ hCD298^+ ; $18.2 \pm 8.7\%$ mMHC I^+ ; $0.5 \pm 0.3\%$ $\text{hCD298}^-\text{mMHC I}^+$; HCI-002: $92.8 \pm 3.2\%$ hCD298^+ ; $5.8 \pm 4.0\%$ mMHC I^+ ; $0.3 \pm 0.2\%$ $\text{hCD298}^-\text{mMHC I}^+$; HCI-010: $97.1 \pm 1.0\%$ hCD298^+ ; $2.0 \pm 0.6\%$ mMHC I^+ ; $0.1 \pm 0.1\%$ $\text{hCD298}^-\text{mMHC I}^+$. In single-cell multiplex qPCR experiments where the number of metastatic cells identified was listed (Extended Data Fig. 2e, #Cells), the entire tissue sample was run through the flow cytometer. A consistent number of live cells was found in tissues from each animal. In any case where live cell yields deviated from the average by more than one standard deviation, mice were excluded from the study (Supplementary Data 8 shows histograms for cell yields from lung and lymph nodes). In Extended Data Table 1 and Fig. 4e, animals or tissues were designated as positive for metastatic cells if $>10 \text{ hCD298}^+\text{mLin}^-$ cells were detected in the entire sample.

Fluidigm dynamic array experiments. Single-cell gene-expression experiments were performed using Fluidigm's 96.96 qPCR DynamicArray microfluidic chips. Single cells were sorted by FACS into individual wells of 96-well PCR plates, using the FACSAriaII single-cell sorting protocol with specific adjustments (device: 96-well PCR plate; precision: single-cell; nozzle: $100 \mu\text{m}$). Experiments were performed according to Fluidigm's Advanced Development Protocol 41. Each well of 96-well PCR plates was preloaded with $9 \mu\text{l}$ volume of RT-STA solution: $5 \mu\text{l}$ of CellsDirect PCR mix (Invitrogen), $0.2 \mu\text{l}$ of SuperScript-III RT/Platinum Taq mix (Invitrogen), $1.0 \mu\text{l}$ of a mixture of all pooled primer assays (500 nM), and $2.8 \mu\text{l}$ of DNA suspension buffer (TEKNOVA). After sorting, PCR plates were frozen (-20°C) or placed into a thermocycler for combined reverse transcription (50°C for 15 min, 95°C for 2 min) and target-specific amplification (20 cycles; each cycle: 95°C for 15 s, 58°C for 4 min). Technical replicates were not performed, as the manufacturer recommends a greater number of biological replicates in lieu of technical replicates yields more power and better sampling of the target population. $3.6 \mu\text{l}$ of exonuclease reaction solution ($2.52 \mu\text{l H}_2\text{O}$, 0.36 Exo reaction buffer, and $0.72 \mu\text{l ExoI}$, New England BioLabs) was then added to remove unincorporated primers (37°C for 30 min, 80°C for 15 min). Subsequently, each well was diluted 1:3 with TE buffer (TEKNOVA). In a separate plate, a $2.7 \mu\text{l}$ aliquot from each sample well was then mixed with $2.5 \mu\text{l}$ of SsoFast EvaGreen Supermix with Low Rox (Bio-Rad) and $0.25 \mu\text{l}$ of Fluidigm's DNA Binding Dye Sample Loading Reagent. Plates were centrifuged to mix solutions. In another separate plate, individual primer assay mixes were generated by loading $2.5 \mu\text{l}$ of Assay Loading Reagent (Fluidigm), $2.25 \mu\text{l}$ DNA Suspension Buffer, and $0.25 \mu\text{l}$ of $100 \mu\text{M}$ primer pair mix. Before loading primer assays and sample mixes into each chip, chips were primed by injecting control line fluid (Fluidigm) and running the 'Prime' program in the IFC Controller HX. After priming, $5 \mu\text{l}$ of each sample and primer mix were loaded into each well of the chips. Samples and assays were then mixed in the chip by running the 'Load Mix' program in the IFC

Controller HX. Chips were transferred into the BioMark real-time PCR reader (Fluidigm) and run according to the manufacturer's instructions. A list of primer assays used in this study is provided in Supplementary Table 1. All primer sequences were acquired through the Harvard Primer bank, and synthesized by Integrated DNA Technologies. Primer assays were run on Fluidigm's dynamic arrays using an iterative approach, where genes that were not informative were replaced in subsequent experiments. Thorough technical evaluations of the microfluidics array technology, limits of detection, and efficiency of multiplex PCR in this platform have been reported by Fluidigm and several independent reports^{29–31}.

Computational analysis, display, and statistical assessment of single-cell PCR data sets. All single-cell PCR data were analysed using Fluidigm's Real-time PCR analysis software, using the Linear (Derivative) and User (Detectors) settings to generate Ct values for each gene. Ct values were further processed in the R statistical language³², using algorithms we generated. All code is provided in Supplementary Information, and published in GitHub (<https://github.com/>) for upload into R. Single-cell multiplex qPCR data are available at the NCBI GEO database (accession GSE70555). Over 20 mice were analysed, but data from 12 PDX mice are included (in which a similar gene set was analysed). Mammary epithelial cells from three reduction mastectomy patients were also analysed. In total, 268 mammary cells from reduction mastectomies, and 441 metastatic and 523 primary tumour cells from PDX mice were analysed.

In normal mammary cell experiments, Ct values were normalized by subtracting the average value of the basal/stem-cell population on a per-gene, per-array basis to correct for batch-to-batch differences in reverse transcription, pre-amplification, and real-time PCR. In PDX experiments, Ct values were normalized by subtracting the average primary tumour expression from the same individual animal on a per-gene basis, to identify conserved differences in gene expression in metastatic cells relative to the primary tumour cells they derive from, in addition to correction of batch-to-batch differences. Normalization using housekeeping genes was not performed, as it is not recommended for single-cell qPCR³³. Normalized Ct values were converted to relative \log_2 expression values simply through multiplication by -1 . Low-quality samples were identified and removed from further analysis in most experiments if less than 80% of the assayed genes amplified. Gene expression data were displayed by PCA, unsupervised hierarchical clustering, supervised clustering, and box plots. Unsupervised hierarchical clustering was performed on both metastatic cells and genes based on Pearson's correlation distance metric and average linkage, after z -score standardization of the \log_2 expression values for each gene across all samples (Fig. 3c). In all other PDX heatmaps, genes were not clustered, but instead the gene order was maintained for consistency. For all heatmaps, the limits of the blue/red colour scale are set to span 90% of the data based on a normal distribution, to prevent outliers from compressing the colours of the majority of the data. For PCA, in which missing data are not easily accommodated, a lower limit of detection approach was taken, in which failed reactions were set to a \log_2 expression value one lower than the minimum observed value across all samples for each gene separately.

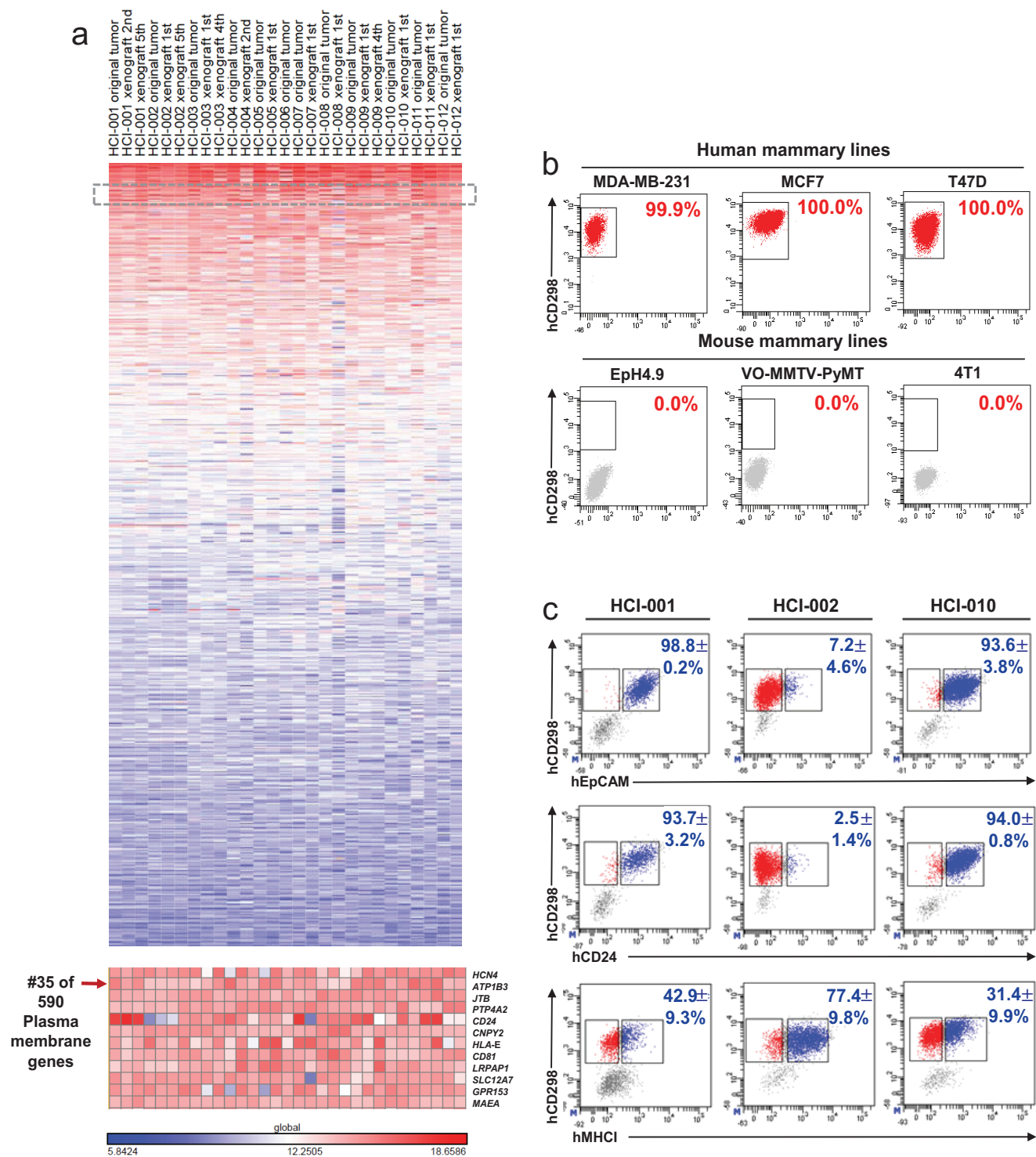
To identify gene expression differences between predefined populations, several statistical tests were performed. For normal mammary cell experiments, we first performed three-group comparisons between basal/stem, luminal, and luminal progenitor cells (both parametric: analysis of variance (ANOVA); and non-parametric: Kruskal–Wallis). This yielded a list of 49 differentially expressed genes (Fig. 1c and Supplementary Table 2). To determine which genes were characteristic of each population, we subsequently performed pair-wise tests (parametric: moderated t -test; and non-parametric: Mann–Whitney U test). In metastatic cell versus primary tumour cell experiments, only pair-wise comparisons were performed. Three-group comparisons were performed to compare lung metastatic cells from the three PDX models, and five-group comparisons were performed to compare metastatic cells from each tissue. In these analyses, ANOVA and Kruskal–Wallis group tests were performed followed by post-hoc pairwise analyses using Tukey and Chi-squared tests. In low- versus high-burden metastatic cell comparisons, low burden was defined as <250 human cells detected in the entire tissue, and high burden was defined as >1,000 cells. Intermediate burden was defined as in between 250 and 1,000 human cells detected. As we were only analysing assays for which at least one cell yielded amplification, undetectable amplification represented non-expression rather than technical error in the PCR reaction. To capture non-expression in the

statistical tests, failed reactions were set to a value 0.01 lower than the lowest observed value across all samples for each gene separately. For the non-parametric tests described earlier, the specific value chosen is not important, while for the parametric tests, this method is comparable to using a lower limit of detection. Our algorithm selected the most appropriate test from which to report a P value based on the type of data observed for that gene (non-parametric if >50% of samples failed for either group, parametric otherwise). This criterion was chosen in an attempt to prevent a high proportion of failed values masking group differences. All P values were also adjusted for the fact that many genes were being simultaneously analysed by controlling the false discovery rate (FDR) with the Benjamini–Hochberg method. To identify basal/stem-cell-characteristic genes, we compared basal/stem (B) to both luminal (L) and luminal progenitor cells (LP) (that is, B versus (L and LP)). Luminal genes were identified by performing L versus B, and luminal progenitor genes by performing LP versus L (since they are a subset of the L lineage). Log-fold changes were computed as a difference between the mean of the \log_2 -normalized expression values for one group versus the mean of the values for the other group; failed reactions were first replaced using the lower limit of detection approach described above.

Enrichment analysis of Biological Process gene ontology terms was performed using the GOstats R package, using the conditional parameter. This was done to identify pathways that were more represented in the set of significantly differentially expressed genes than would be expected by chance alone.

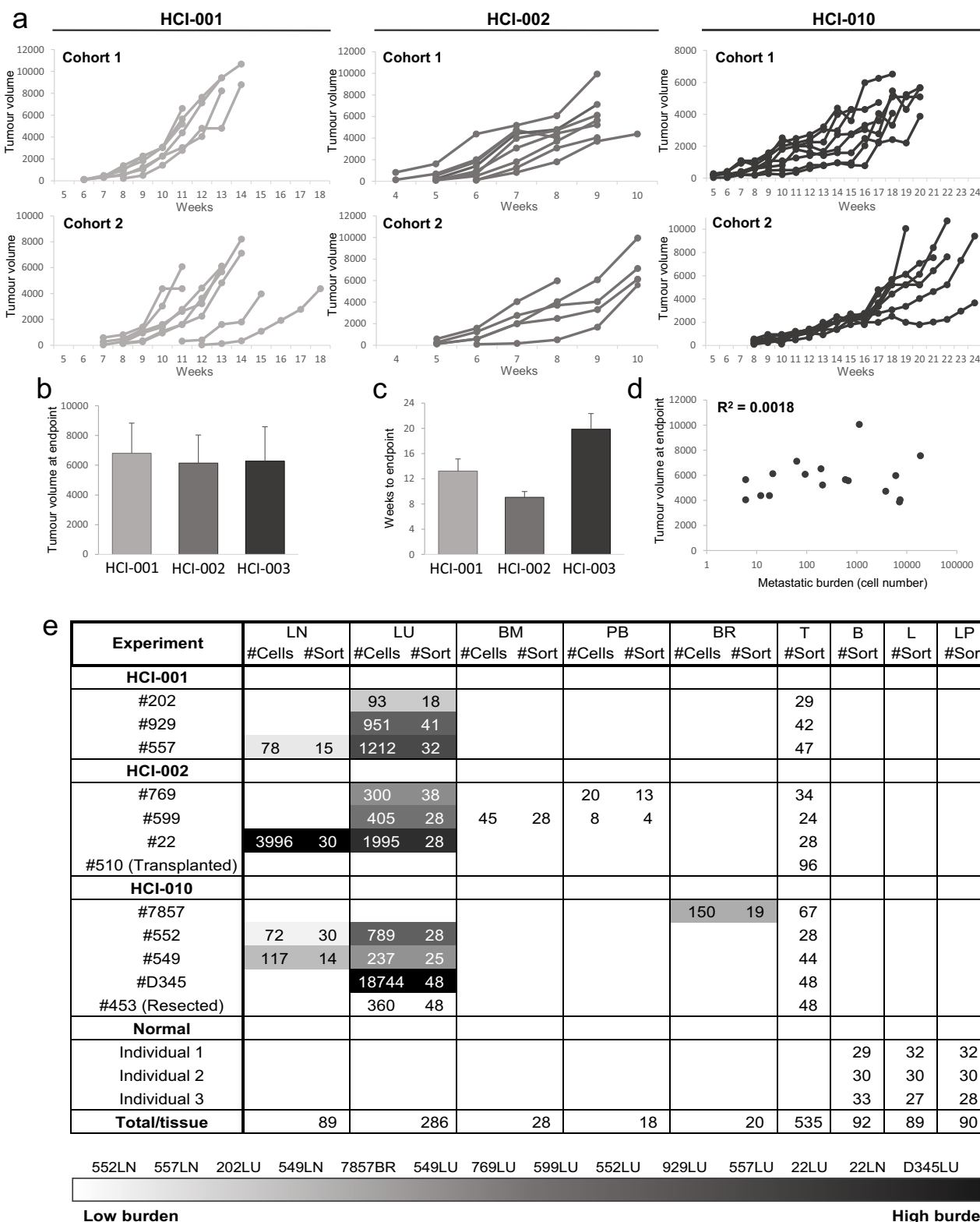
Histological and immunofluorescent analysis. Tissues were fixed overnight in 4% paraformaldehyde and processed for paraffin embedding. For histological analysis, sections were stained with haematoxylin and eosin using standard methods. Immunofluorescent staining was performed on lung tissues with low and high metastatic burden. We defined low burden as fewer than 10 small detectable lesions, containing fewer than 20 cells each. High burden was defined as greater than 25 lesions, with large numbers of cells (at least 1,000 in total). Metastatic lesions were identified by the size of the nuclei, as tumour cell nuclei were 2–3 times larger than surrounding nuclei in the lung. Metastatic lesions were also often encircled by basement membrane and stroma, making them easily identifiable. Immunostaining on paraffin-embedded tissue sections was performed using standard protocols, using citrate buffer (pH 6.0) and heating in a pressure cooker for 8 min. MUC1 (Sigma, HPA008855, 1:100) and KRT5 (Biolegend, PRB-160P, 1:1,000) were stained using a three-step technique, where primary antibodies were incubated overnight, followed by 1 h incubations with a biotinylated anti-rabbit secondary (DAKO, 1:500) and subsequently a Streptavidin Alexa-568 (Invitrogen, 1:1,000). MYC (abcam, ab32072, 1:100) and phospho-histone H3 (Cell Signaling Technology, 1:100) were identified using a two-step technique, where overnight primary antibody stains were followed by a 1 h incubation with a goat anti-rabbit Alexa-568 secondary (Molecular Probes, 1:1,000). The number of positive nuclei was counted in several fields for each group (tumour, high burden, and low burden), and significance was calculated by single-factor ANOVA and pair-wise t -tests assuming equal variance.

25. Parry, D. *et al.* Dinaciclib (SCH 727965), a novel and potent cyclin-dependent kinase inhibitor. *Mol. Cancer Ther.* **9**, 2344–2353 (2010).
26. Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proc. Natl Acad. Sci. USA* **105**, 20380–20385 (2008).
27. Györfy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* **123**, 725–731 (2010).
28. Nguyen-Ngoc, K. V. *et al.* ECM microenvironment regulates collective migration and local dissemination in normal and malignant mammary epithelium. *Proc. Natl Acad. Sci. USA* **109**, E2595–E2604 (2012).
29. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnol.* **29**, 1120–1127 (2011).
30. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
31. Devonshire, A. S., Elawarapu, R. & Foy, C. A. Applicability of RNA standards for evaluating RT-qPCR assays and platforms. *BMC Genomics* **12**, 118 (2011).
32. R. Development Core Team. *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2012).
33. McDavid, A. *et al.* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461–467 (2013).



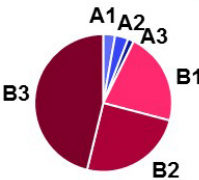
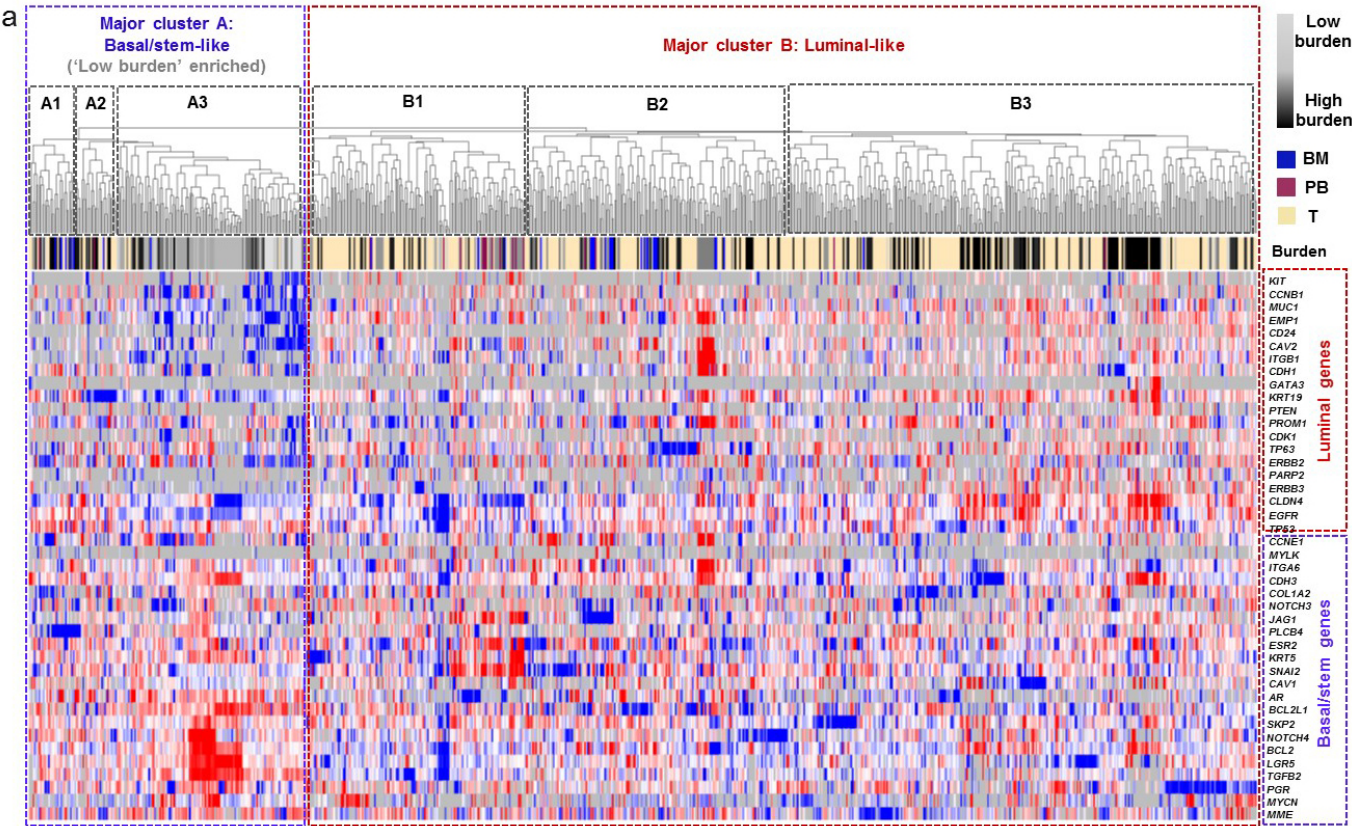
Extended Data Figure 1 | Identification and validation of CD298 for detection of human cells. **a**, Analysis of published microarray data identified *CD298* as highly expressed on many PDX breast cancer models and corresponding original patient tumours. The heatmap shows genes rank ordered from highest to lowest for raw expression values across all samples. The inset (bottom) highlights expression for *CD298* (also known as *ATP1B3*). *CD298* ranked number 35 out of over 590 plasma membrane genes. **b**, FACS for

CD298 on human (top) and mouse (bottom) mammary cell lines to establish species specificity. **c**, FACS on primary PDX tumour cells comparing *CD298* expression with other markers used in related applications (EpCAM, CD24, MHC I; percentages indicate dual-positive cells) ($n = 3$). EpCAM is used to identify CTCs in the clinic; CD24 is a pan-epithelial marker; and MHC I is used as a ubiquitous marker on all nucleated cells. These markers were not used in this study because they were not robustly expressed on all PDX models.

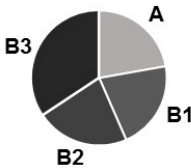
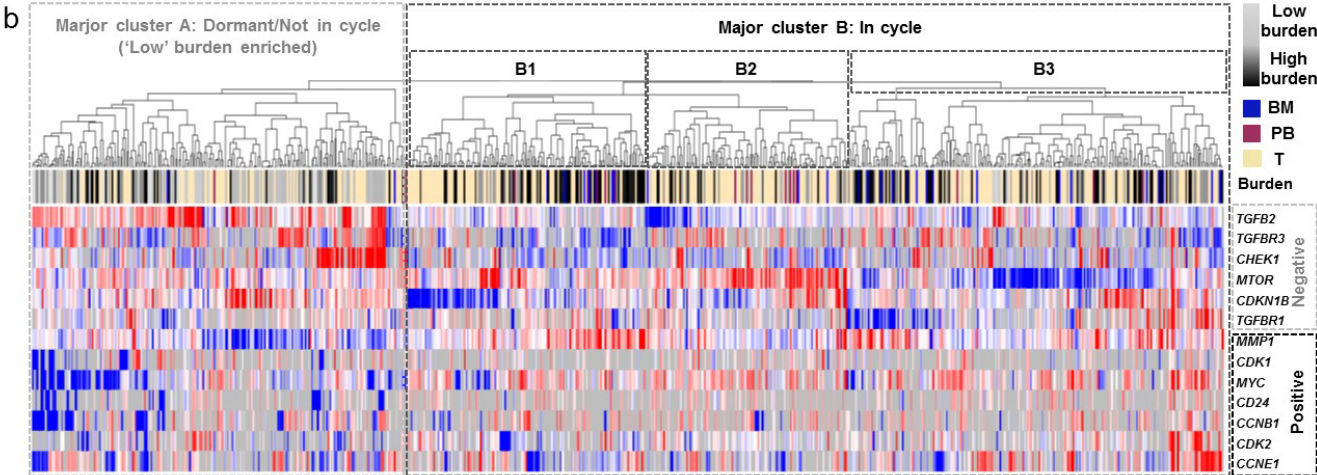


Extended Data Figure 2 | Analysis of primary tumour growth kinetics and metastasis in PDX mice. **a**, Weekly caliper measurements of primary tumours in two independent cohorts of animals show that growth kinetics were consistent within each PDX model. **b**, Bar graph shows that the average tumour volume at the endpoint was similar across PDX models. **c**, Bar graph shows the average number of weeks for tumours to reach endpoint (20–25 mm diameter) in each PDX model. **d**, Correlation plot shows that metastatic burden did not correlate with tumour volume in PDX animals. **e**, Table summarizing the number of metastatic cells detected (#Cells) and analysed (#Sort) from each tissue from each PDX animal. Tissues were rank ordered

according to metastatic burden, from lowest (lightest grey) to highest (black). The table also shows the number of primary tumour cells analysed (#Sort) from PDX animals, and the number of normal mammary epithelial cells analysed (#Sort) from mastoplasia patients (Individuals 1, 2, and 3). 'Transplanted' indicates primary tumour cells derived from transplant of lymph node metastatic cells into fatty fat pads; 'resected' indicates lung metastatic cells analysed 8 weeks after resection of the primary tumour. B, basal/stem; BM, bone marrow; BR, brain; L, luminal; LN, lymph node; LP, luminal progenitor; LU, lung; PB, peripheral blood; T, primary tumour.



| | | Basal/stem-like | | | | Luminal-like | | |
|---------|-----------|-----------------|------|------|------|--------------|-------|-------|
| | Met freq. | Total A | A1 | A2 | A3 | Total B | B1 | B2 |
| Total | | 7.3% | 2.8% | 3.1% | 1.4% | 92.7% | 21.8% | 24.7% |
| HCI-001 | Intermed. | 6.0% | 2.6% | 3.4% | 0.0% | 94.0% | 23.1% | 35.0% |
| HCI-002 | Lowest | 3.3% | 2.2% | 0.0% | 1.1% | 96.7% | 25.9% | 15.7% |
| HCI-010 | Highest | 11.6% | 3.4% | 5.5% | 2.7% | 88.4% | 19.9% | 21.9% |

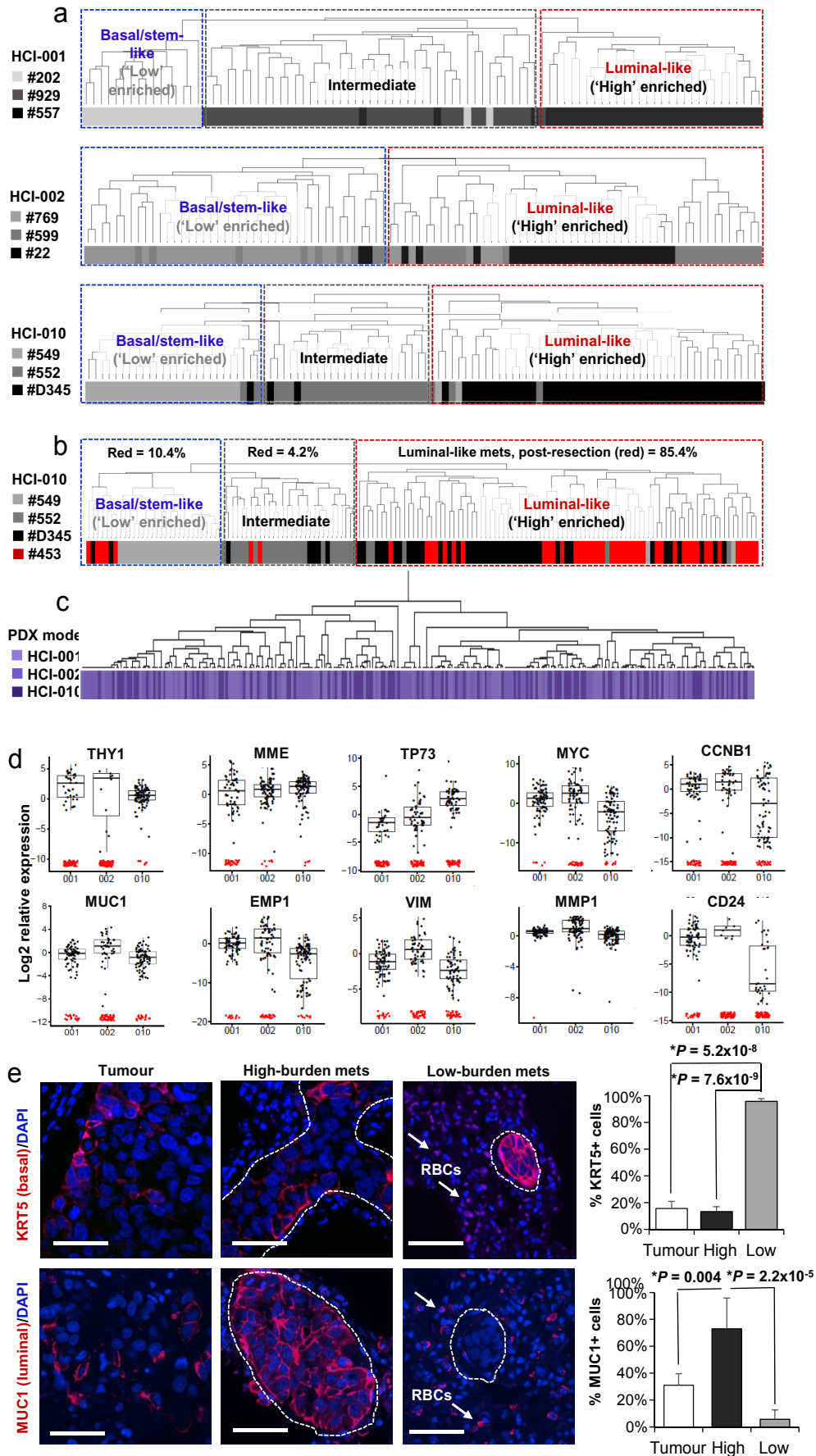


| | | Not in cycle | | In cycle | | |
|---------|-----------|--------------|---------|----------|-------|-------|
| | Met freq. | Total A | Total B | B1 | B2 | B3 |
| Total | | 22.2% | 77.8% | 21.2% | 22.2% | 34.4% |
| HCI-001 | Intermed. | 21.5% | 78.5% | 21.4% | 23.8% | 33.3% |
| HCI-002 | Lowest | 35.2% | 64.8% | 31.0% | 14.1% | 19.7% |
| HCI-010 | Highest | 23.0% | 77.0% | 20.8% | 24.7% | 31.5% |

Extended Data Figure 3 | Primary tumours contain rare stem-like cells.

a, Unsupervised hierarchical clustering of metastatic and primary tumour cells from 10 animals (Extended Data Fig. 2e lists cells analysed from each animal) based on their expression of the 49-gene differentiation signature. The dendrogram shows two major clusters, where major cluster A contains basal/stem-like cells and major cluster B contains more luminal-like cells. The majority of low-burden metastatic cells reside in subcluster A3. 1.4% of the primary tumour cells analysed in this study reside in subcluster A3, and are therefore similar to low-burden metastatic cells in their stem-like differentiation status. The pie graph and table list the percentage of primary

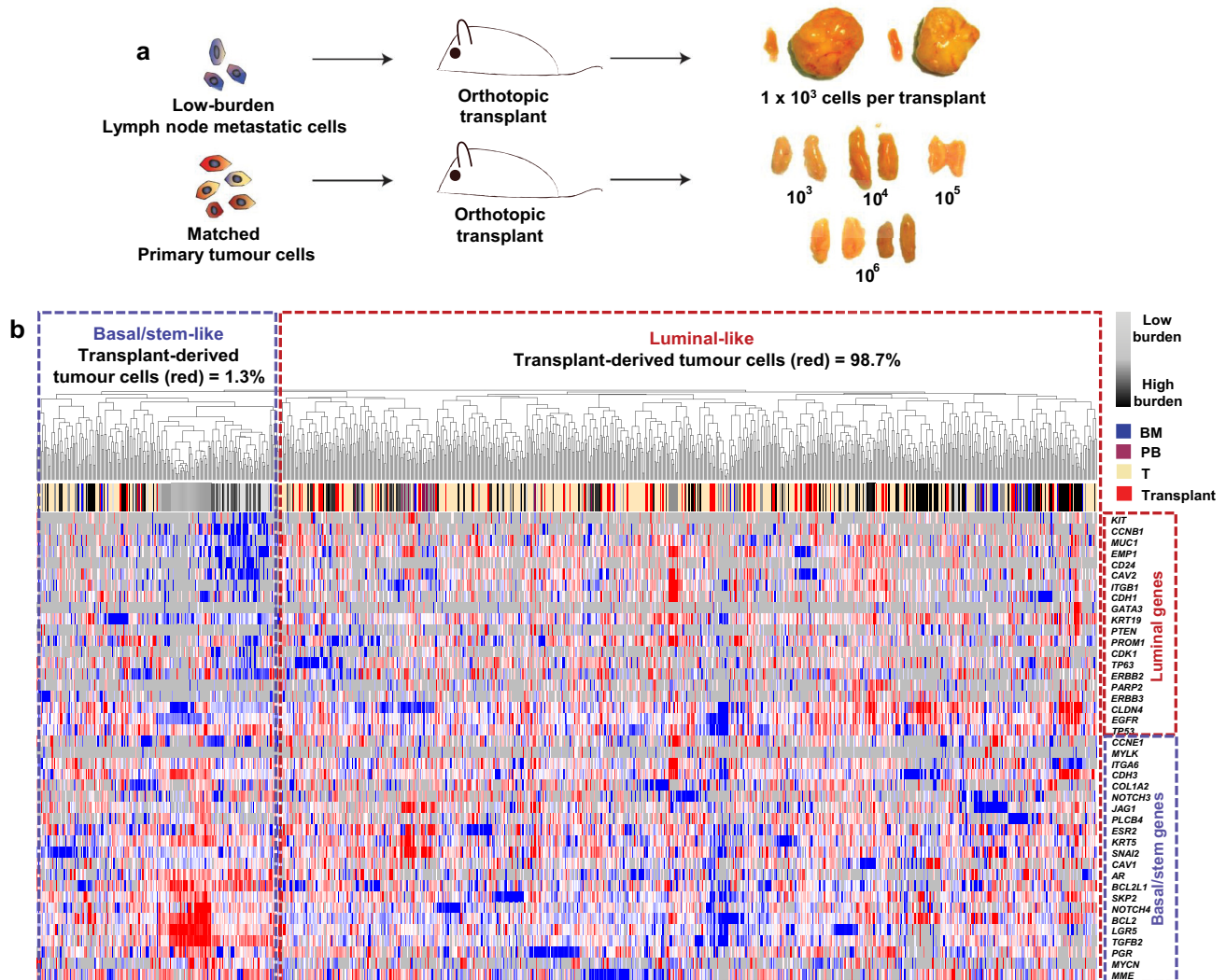
tumour cells that reside in each cluster. The table also shows the data by PDX model. **b**, Unsupervised hierarchical clustering of metastatic and primary tumour cells, based on their expression of genes associated with cell cycle and dormancy. Two major clusters are evident. Major cluster A contains cells with a less-proliferative signature, which express higher levels of 'negative' cell-cycle-associated genes and lower levels of 'positive' cell-cycle-associated genes. Major cluster B contains cells with a more proliferative signature. The majority of low-burden metastatic cells reside in major cluster A, and possess a less-proliferative signature. The pie chart and table show the number of primary tumour cells in each cluster.



Extended Data Figure 4 | The correlation between differentiation and metastatic burden is conserved in each PDX model.

a, Unsupervised hierarchical clustering of lung metastatic cells from each PDX model is shown separately. Lung metastatic cells were specifically chosen for this analysis because they were the only tissue for which there were sufficient numbers of low- and high-burden cells. In each dendrogram, low-burden metastatic cells form a distinct cluster due to their basal/stem-like expression signature. High-burden metastatic cells also form distinct clusters and express higher levels of luminal genes. Supplementary Data 2 shows the entire heatmap for each PDX model. **b**, Unsupervised hierarchical clustering of lung metastatic cells that developed after primary tumour resection (#453, red) at 10–12 mm in diameter. Post-resection metastatic cells were clustered with lung metastatic cells from non-resected animals to investigate their differentiation status. All animals bore the HCI-010 model. 85.4% of post-resection metastatic cells displayed a luminal-like expression pattern, showing that luminal-like metastatic cells can arise from cells that disseminate at early stages of primary tumour growth. Supplementary Data 2 shows the entire heatmap. **c**, Unsupervised hierarchical

clustering of lung metastatic cells from all three PDX models by their expression of the top genes differentially expressed between them. Although there were statistically significant differences between the models, the dendrogram shows that they were not powerful enough to cluster the cells separately by model. Supplementary Data 2 shows the entire heatmap. **d**, Box plots show top selected genes differentially expressed between the three PDX models. By ANOVA, 53 genes were significantly differentially expressed ($P < 0.05$, Supplementary Data 3). **e**, Immunofluorescence stains for basal and luminal lineage-specific proteins (red) in micro- and macrometastatic lesions. Autofluorescent red blood cells (RBCs) are also present in the lung (arrows), but do not represent positive immunostaining. Scale bars, 50 μm . Bar graphs quantify the percentage of low- and high-burden metastatic cells, and primary tumour cells that were positive for antibody staining. Data from at least three fields, in three different mice was collected from each group, and P values were calculated as described in the Methods. Error bars represent standard deviation.



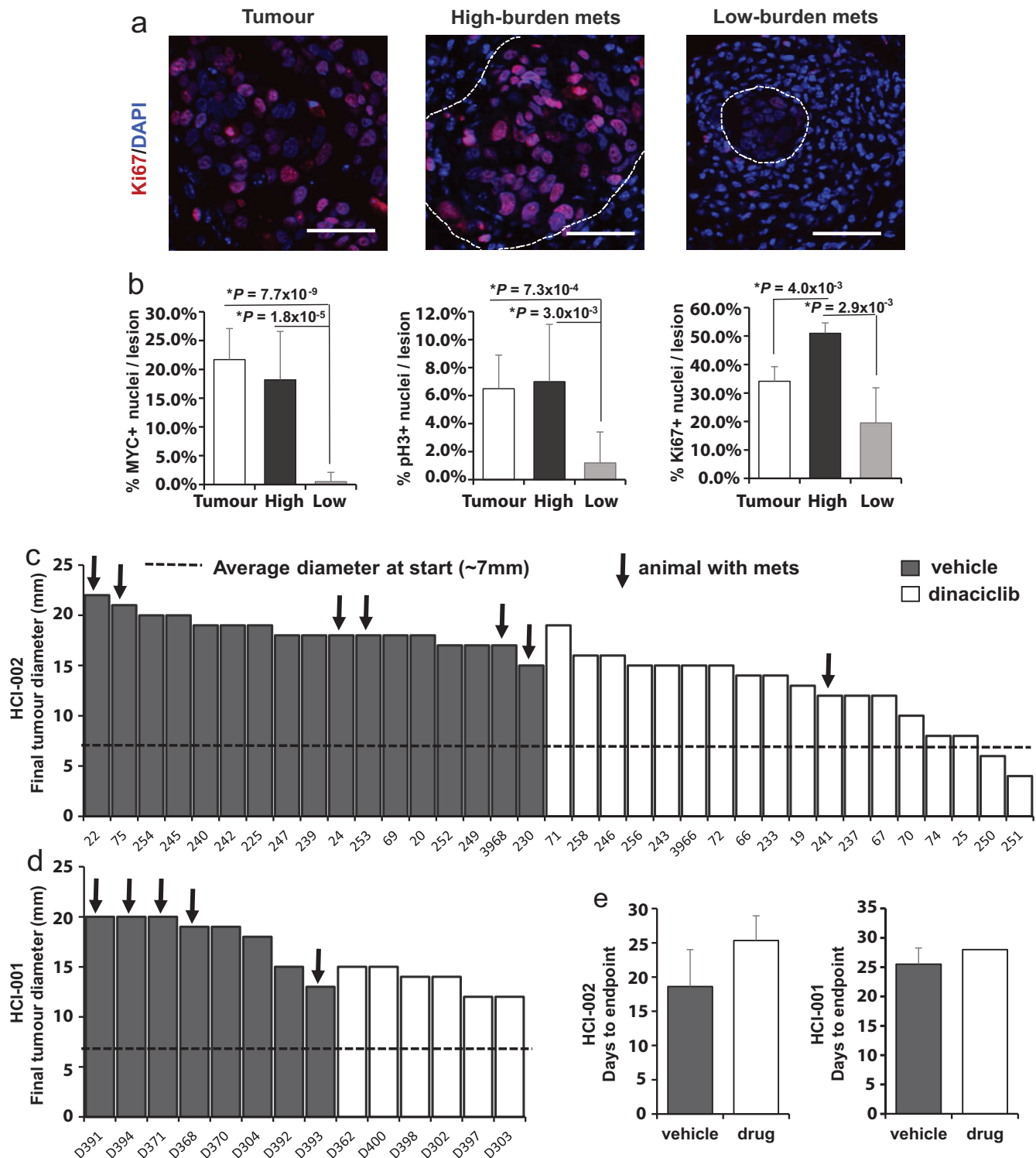
Extended Data Figure 5 | Low-burden metastatic cells have tumour-initiating and differentiation capacity. **a**, Schematic overview of orthotopic transplant experiments to investigate the tumour-initiating and differentiation capacity of low-burden metastatic cells. Images of resulting grafts show that 2/4 transplants of low-burden cells grew large tumours, while 0/10 transplants from primary tumour cells developed tumours. **b**, Unsupervised hierarchical clustering of tumour cells derived from transplants of low-burden metastatic

cells. Transplant-derived tumour cells were clustered with metastatic and primary tumour cells from previous experiments (Extended Data Fig. 3a) to investigate their differentiation status. Transplant-derived tumour cells were heterogeneous, where 1.3% of them were basal/stem-like, and 98.7% of them clustered with more luminal-like cells. This shows that low-burden basal/stem-like metastatic cells have the capacity to give rise to luminal-like cancer cells.



Extended Data Figure 6 | Metastatic cells found in different organs show distinct gene expression signatures. **a**, Supervised clustering of metastatic cells by target organ emphasizes tissue-specific gene signatures. Arrows indicate genes significantly differentially expressed between at least two tissues, as shown in **b**. **b**, Box plots show genes most characteristic of each tissue type, as determined by ANOVA and pair-wise analyses. *P* values and fold change for each gene and tissue pair are listed in Supplementary Table 3. Box plots for all

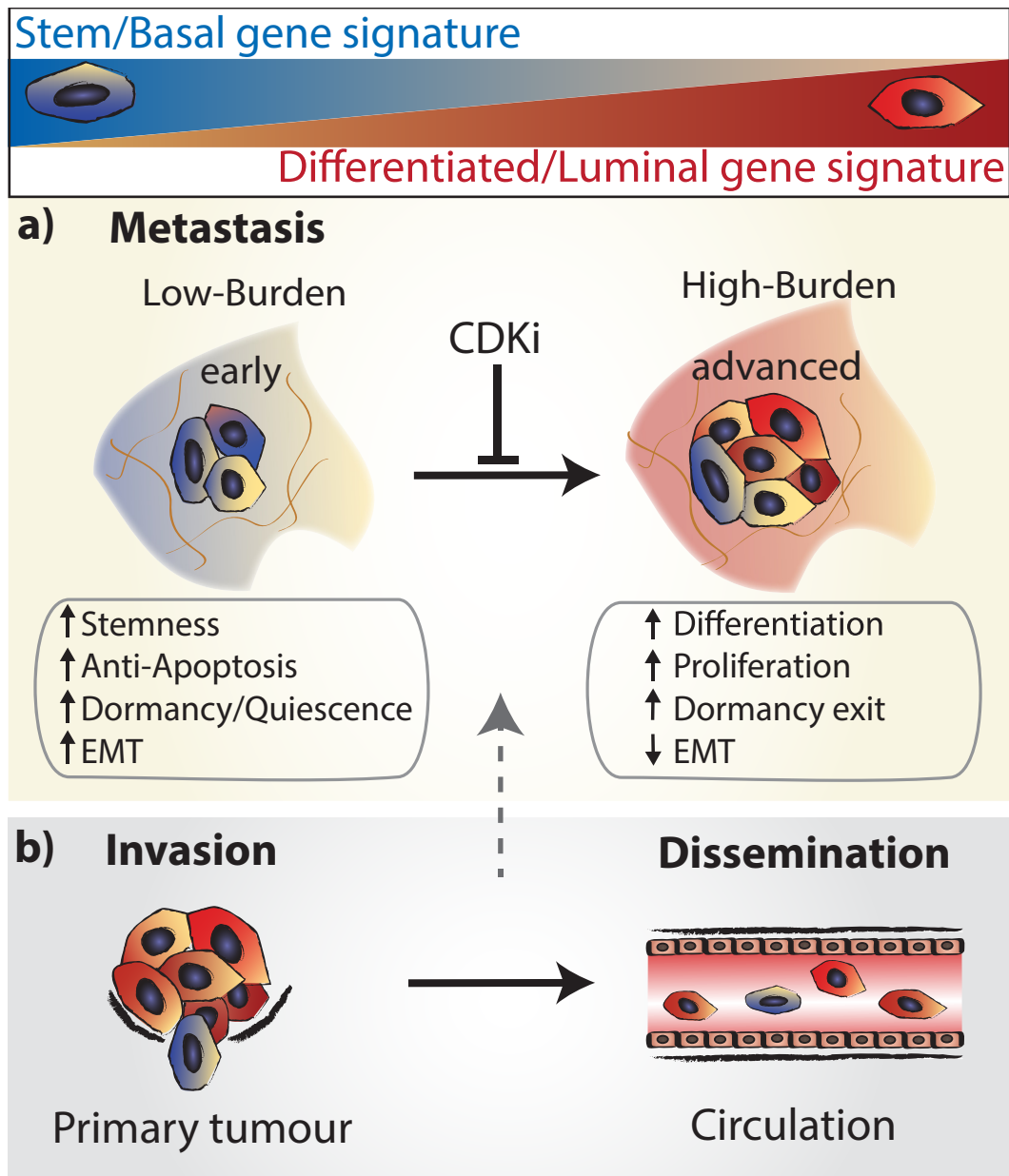
80 genes differentially expressed between the tissue pairs are shown in Supplementary Data 7. BM, bone marrow; BR, brain; LN, lymph node; LU, lung; PB, peripheral blood (CTC); T, tumour. **c**, Pearson correlations indicate similarity of CTCs to other metastatic tissue types across all genes analysed. Each dot represents an individual gene. BM, bone marrow; BR, brain; LN, lymph node; LU, lung; PB, peripheral blood (CTC).



Extended Data Figure 7 | Analysis of dinaciclib-treated animals.

a, Immunofluorescence stains for Ki67 in micro- and macrometastatic lesions from low- and high-burden animals, as well as in primary tumours. Scale bars, 50 µm. **b**, Bar graphs quantify the percentage of MYC, phospho-histone H3 (pH3), and Ki67 positive cells per lesion in micro- and macrometastatic lesions.

Error bars represent standard deviation. **c**, **d**, Waterfall plots shows the longest final tumour diameter for each PDX animal treated with vehicle (black bars) or drug (white bars). **e**, Bar graphs show the average number of days to endpoint (4 weeks, or 20 mm primary tumour size) for animals treated with vehicle or drug.



Extended Data Figure 8 | Model for tumour cell heterogeneity during metastatic progression. **a**, Metastatic cells from animals with low metastatic burden (blue) are distinct from animals with higher burden, due to their increased expression of stemness, anti-apoptosis, EMT, and dormancy/quiescence-related genes. In contrast, higher burden metastatic cells are more heterogeneous, and comprise larger numbers of proliferative, differentiated cells (red). Transplant experiments of stem-like metastatic cells showed that they have tumour-initiating potential, and can produce luminal-like cancer cells. This strongly suggests that metastases derive from stem-like cells, which differentiate and undergo a switch from dormancy into proliferation as they

colonize and produce more advanced metastatic tumours. Metastatic progression was also associated with increased MYC expression, and could be attenuated with CDK inhibition. We believe this is due to apoptosis of cells as they upregulate MYC, since our previous work has shown that CDK inhibition induces apoptosis in high MYC-expressing cancer cells through synthetic lethality²³. **b**, Comparison of gene signatures in primary tumour and metastatic cells showed that 1.4% of primary tumour cells, and 16.7% of CTCs possessed a stem-like signature. This suggests that these cells may be the origin of metastatic tumours.

Extended Data Table 1 | Metastatic frequency and tissue tropism identified by FACS in each PDX model

| | HCI-001 | HCI-002 | HCI-010 |
|-----------------------|---------------|---------------|------------------|
| PATIENT | | | |
| Marker status | ER-/PR-/Her2- | ER-/PR-/Her2- | ER-/PR-/Her2- |
| Tumor subtype (PAM50) | Basal | Basal | Basal |
| Sample source | Breast | Breast | Pleural effusion |
| Diagnosed metastases | Lymph node | Lymph node | Lung |
| PDX MICE | | | |
| Total mice with mets | 26/32 (81%) | 13/35 (37%) | 31/33 (94%) |
| Peripheral blood | 7/31 | 1/19 | 3/22 |
| Lymph node | 8/31 | 7/34 | 17/33 |
| Lung | 26/32 | 7/35 | 31/33 |
| Bone marrow | 3/21 | 4/25 | 6/23 |
| Brain | 0/15 | 1/8 | 1/13 |

Extended Data Table 2 | All genes differentially expressed in low-burden metastatic cells relative to primary tumour cells

| Increased expression | | | | Decreased expression | | | |
|----------------------|----------------|----------------------------|-----------------------|----------------------|----------------|----------------------------|------------------------|
| Gene | Normal lineage | Fold change (low-burden/T) | *P value | Gene | Normal lineage | Fold change (low-burden/T) | *P value |
| TGFBR2 | N/A | 43.0 | 4.8x10 ⁻¹² | PTEN | luminal | -2.1 | 0.001 |
| BCL2L1 | basal/stem | 30.1 | 2.2x10 ⁻¹¹ | TGFBR1 | N/A | -2.3 | 0.028 |
| EPHA4 | N/A | 27.0 | 4.5x10 ⁻¹⁴ | ERBB3 | luminal | -2.3 | 0.004 |
| AR | luminal | 21.5 | 7.3x10 ⁻¹⁵ | CDH1 | luminal | -2.6 | 0.059 |
| LGR5 | basal/stem | 17.6 | 4.6x10 ⁻¹³ | CDK2 | N/A | -3.5 | 0.001 |
| IGFBP6 | N/A | 15.7 | 2.4x10 ⁻⁸ | MUC1 | luminal | -3.9 | 0.007 |
| TGFB2 | basal/stem | 15.7 | 2.4x10 ⁻¹⁵ | VEGFA | N/A | -6.6 | 3.2x 10 ⁻⁵ |
| SOX2 | N/A | 14.5 | 2.8x10 ⁻¹¹ | CAV2 | basal/stem | -7.5 | 2.3x10 ⁻⁵ |
| BMI1 | basal/stem | 14.0 | 1.1x10 ⁻²⁸ | MYC | N/A | -9.8 | 2.3x10 ⁻⁵ |
| CXCL12 | N/A | 13.7 | 8.1x10 ⁻⁶ | ITGB1 | basal/stem | -11.2 | 9.1 x10 ⁻²⁶ |
| TWIST1 | N/A | 13.4 | 2.5x10 ⁻⁷ | PARP2 | luminal prog. | -15.8 | 8.3x10 ⁻¹¹ |
| BCL2 | basal/stem | 12.3 | 1.3x10 ⁻¹⁰ | EMP1 | luminal | -20.2 | 3.3x10 ⁻⁷ |
| NOTCH4 | basal/stem | 11.6 | 1.0x10 ⁻¹⁶ | CD24 | luminal | -28.7 | 1.7x10 ⁻¹² |
| KRT5 | basal/stem | 10.8 | 6.1x10 ⁻⁸ | VIM | N/A | -29.8 | 4.4x10 ⁻¹⁴ |
| POU5F1 | N/A | 10.5 | 1.8x10 ⁻⁶ | | | | |
| TGFB1 | N/A | 8.1 | 7.1x10 ⁻⁵ | | | | |
| THY1 | N/A | 7.4 | 2.3x10 ⁻⁵ | | | | |
| CDKN1B | N/A | 7.3 | 5.1x10 ⁻¹¹ | | | | |
| WNT2 | N/A | 6.6 | 6.7x10 ⁻⁶ | | | | |
| SKP2 | basal/stem | 6.2 | 4.7x10 ⁻⁷ | | | | |
| DAND5 | N/A | 6.1 | 2.2x10 ⁻⁹ | | | | |
| PGR | basal/stem | 5.8 | 9.4x10 ⁻⁵ | | | | |
| CHEK1 | basal/stem | 5.6 | 2.3x10 ⁻⁵ | | | | |
| CDH3 | basal/stem | 4.3 | 1.5x10 ⁻⁹ | | | | |
| MTOR | basal/stem | 3.8 | 1.0x10 ⁻⁴ | | | | |
| TP73 | N/A | 3.5 | 5.1x10 ⁻⁷ | | | | |
| TGFBR3 | N/A | 3.4 | 0.004 | | | | |
| ESR2 | basal/stem | 3.3 | 0.002 | | | | |
| ESR1 | N/A | 3.3 | 0.004 | | | | |
| MAX | N/A | 3.2 | 0.010 | | | | |
| NTRK2 | N/A | 2.9 | 0.014 | | | | |
| NOTCH3 | basal/stem | 2.9 | 0.040 | | | | |
| FIGF | N/A | 2.9 | 0.004 | | | | |
| MME | basal/stem | 2.4 | 0.017 | | | | |
| TP63 | basal/stem | 2.3 | 0.017 | | | | |
| TP53 | basal/stem | 2.2 | 0.003 | | | | |
| MYCN | basal/stem | 2.2 | 0.033 | | | | |
| SNAI2 | basal/stem | 2.1 | 0.046 | | | | |
| ITGA6 | basal/stem | 2.1 | 0.003 | | | | |
| JAG1 | basal/stem | 1.9 | 0.057 | | | | |
| ACTA2 | basal/stem | 1.8 | 0.004 | | | | |

N/A, not part of 49-gene signature; that is, not differentially expressed in normal mammary lineages. P values: moderated *t*-test or Mann-Whitney U test, FDR corrected.

Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins

Joseph Bondy-Denomy^{1†}, Bianca Garcia¹, Scott Strum², Mingjian Du¹, MaryClare F. Rollins³, Yurima Hidalgo-Reyes¹, Blake Wiedenheft³, Karen L. Maxwell⁴ & Alan R. Davidson^{1,2}

The battle for survival between bacteria and the viruses that infect them (phages) has led to the evolution of many bacterial defence systems and phage-encoded antagonists of these systems. Clustered regularly interspaced short palindromic repeats (CRISPR) and the CRISPR-associated (cas) genes comprise an adaptive immune system that is one of the most widespread means by which bacteria defend themselves against phages^{1–3}. We identified the first examples of proteins produced by phages that inhibit a CRISPR–Cas system⁴. Here we performed biochemical and *in vivo* investigations of three of these anti-CRISPR proteins, and show that each inhibits CRISPR–Cas activity through a distinct mechanism. Two block the DNA-binding activity of the CRISPR–Cas complex, yet do this by interacting with different protein subunits, and using steric or non-steric modes of inhibition. The third anti-CRISPR protein operates by binding to the Cas3 helicase–nuclease and preventing its recruitment to the DNA-bound CRISPR–Cas complex. *In vivo*, this anti-CRISPR can convert the CRISPR–Cas system into a transcriptional repressor, providing the first example—to our knowledge—of modulation of CRISPR–Cas activity by a protein interactor. The diverse sequences and mechanisms of action of these anti-CRISPR proteins imply an independent evolution, and foreshadow the existence of other means by which proteins may alter CRISPR–Cas function.

CRISPR–Cas RNA-guided immune systems are widespread in prokaryotes, and play a major part in microbial evolution^{2,3}. In these systems, CRISPR arrays are transcribed and processed to generate small CRISPR RNAs (crRNAs), which combine with Cas proteins to form crRNA-guided surveillance complexes^{2,5}. In the type I-F CRISPR–Cas system, the Csy4 protein is a CRISPR-specific endoribonuclease that binds to and cleaves each repeat sequence in the pre-crRNA⁶. Csy4 remains associated with the 3' end of the mature 60-nucleotide crRNA and then assembles with Csy1, Csy2 and Csy3 proteins to form a 350 kilodalton (kDa) surveillance complex^{7,8}. This complex relies on a 32-nucleotide segment of the crRNA for complementary base pairing to invading DNA sequences, known as protospacers. Binding of target DNA by the Csy complex leads to the recruitment of the nuclease–helicase protein Cas3 and subsequent phage genome degradation^{9,10}. We previously identified five unique type I-F anti-CRISPR proteins⁴. Here we determine the mechanisms by which three of these proteins function.

Three type I-F anti-CRISPRs, AcrF1 (11 kDa, encoded by gene 35 from phage JBD30), AcrF2 (13 kDa, encoded by gene 30 from phage D3112), and AcrF3 (16 kDa, encoded by gene 35 from phage JBD5), could be expressed in *Escherichia coli* and purified to homogeneity. Using a previously described *E. coli* expression system⁷, we also purified the 350 kDa *Pseudomonas aeruginosa* Csy complex, including a crRNA and the four Csy proteins. This complex was mixed *in vitro* with each purified anti-CRISPR protein, and fractionated by size-exclusion chromatography (SEC). AcrF1 and AcrF2 co-eluted with the Csy complex (Fig. 1a and Extended Data Fig. 1), indicating a direct

interaction. AcrF3 did not co-elute with the Csy complex (Fig. 1b). The lack of AcrF3 binding to the Csy complex suggested that it might inhibit the CRISPR–Cas system by interacting with Cas3, the helicase–nuclease that is responsible for target DNA destruction after recognition by the Csy complex. Supporting this hypothesis, AcrF3 co-eluted with purified Cas3, while AcrF1 did not (Fig. 1c and Extended Data Fig. 2). These experiments demonstrate that each of the three tested anti-CRISPR proteins can bind to either the Csy complex or Cas3.

The Csy complex recognizes foreign DNA targets through sequential recognition of a protospacer adjacent motif (PAM) and crRNA-guided base pairing to a target¹¹. We performed electrophoretic mobility shift assays (EMSAs) to demonstrate that the interaction of AcrF1 and AcrF2 with the Csy complex blocked its ability to bind a 50 base pair (bp) double-stranded DNA (dsDNA) target containing a PAM and a sequence identical to the crRNA spacer (Fig. 1d). We used isothermal titration calorimetry to show that these anti-CRISPRs also blocked binding of the Csy complex to an 8-nucleotide single-stranded DNA (ssDNA) target complementary to the functionally crucial 'seed' region¹² of the crRNA (Extended Data Fig. 3). AcrF3, which does not interact with the Csy complex, did not inhibit the DNA-binding activity of the Csy complex (Fig. 1d, lane 5, and Extended Data Fig. 3).

To probe the potential role of AcrF3 in blocking Cas3 activity, we mixed purified Cas3 with the Csy complex and target DNA. In this instance, a supershifted species appeared in the EMSA gel that we presumed comprised the Csy complex, DNA and Cas3 (Fig. 1d, lane 7; a reaction containing only Cas3 and DNA did not display this species, lane 6). Importantly, pre-incubation of Cas3 with AcrF3 prevented formation of the supershifted complex (Fig. 1d, lane 10), indicating that this anti-CRISPR blocks recruitment of Cas3 to the Csy–DNA complex. Pre-incubation of Cas3 with AcrF1 or AcrF2 did not have this effect (Fig. 1d, lanes 8, 9). Further corroborating the presence of Cas3 in the supershifted complex, the addition of ATP prevented formation of this species (Fig. 1d, lane 11) and destabilized a preformed complex (lane 13), probably owing to the activation of the ATP-dependent helicase activity of Cas3, as described for the type I-E CRISPR–Cas system¹⁰.

To demonstrate that the described anti-CRISPR mechanisms operate *in vivo*, we targeted the Csy complex to the promoter of the *phzM* gene, which is required in *P. aeruginosa* for production of the blue-green pigment pyocyanin¹³. Binding of the *phzM* promoter by a Csy complex in the absence of Cas3 activity was expected to repress transcription, as was previously observed for a type I-E CRISPR–Cas system^{14,15}. Consistent with this expectation, targeting of the *phzM* promoter in cells containing a prophage expressing *acrF3* resulted in cultures with a complete lack of pigment production, similar to a strain lacking Cas3 (Fig. 2a; the somewhat higher pigment production in the $\Delta cas3$ strain is probably due to reduced Csy function¹⁶). By contrast, the expression of *acrF1* and *acrF2*, which inhibit DNA binding by the Csy complex, resulted in blue-green cultures, as did expression of the *phzM* promoter targeting crRNA in cells lacking Csy3. Quantitative

¹Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. ²Department of Biochemistry, University of Toronto, Toronto, Ontario M5S 1A8, Canada. ³Department of Microbiology and Immunology, Montana State University, Bozeman, Montana 59717, USA. ⁴Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada. [†]Present address: Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, California 94158, USA.

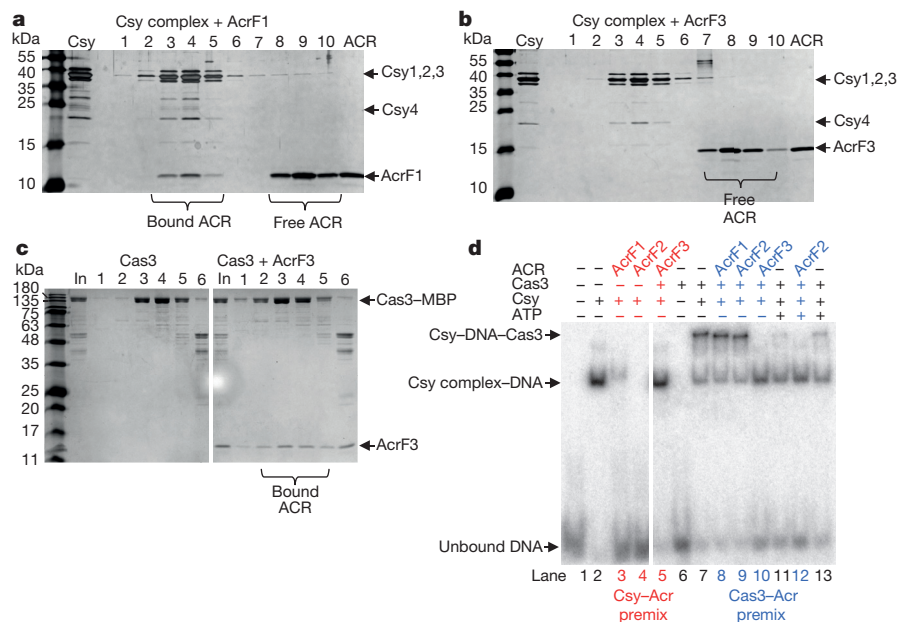


Figure 1 | Anti-CRISPR proteins inhibit CRISPR–Cas function by directly interacting with the Csy complex or Cas3. **a, b,** Purified Csy complex was incubated with purified AcrF1 (**a**) or AcrF3 (**b**) and the mixture was fractionated by SEC. Fractions were analysed by SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and are numbered according to their elution position (see Extended Data Fig. 1 for SEC of the Csy complex alone or with AcrF2). The purified Csy complex or anti-CRISPR (ACR) are shown in the second (Csy) and last (ACR) lanes, respectively. **c,** Purified Cas3 was incubated with (right) or without (left) AcrF3 and fractionated by SEC. The eluting fractions were analysed by SDS–PAGE as described earlier. The input (In) lanes show the protein mixture that was loaded onto the SEC column. MBP,

maltose-binding protein. **d**, dsDNA binding by the Csy complex was assayed using an EMSA. Csy complex was present in all reactions except for lanes 1 and 6. Other components added to each reaction are designated above the lanes. In the lanes coloured red and blue, the designated components were premixed before the addition of DNA. ATP was added to the Csy–DNA–Cas3 reaction either before the addition of Cas3 (lanes 11, 12) or after (lane 13). The supershifted species resulting from Cas3 addition did not migrate into the gel upon prolonged electrophoresis, but it is dissociated by the addition of ATP (lane 13), demonstrating that the supershift is not caused by aggregated inactive protein.

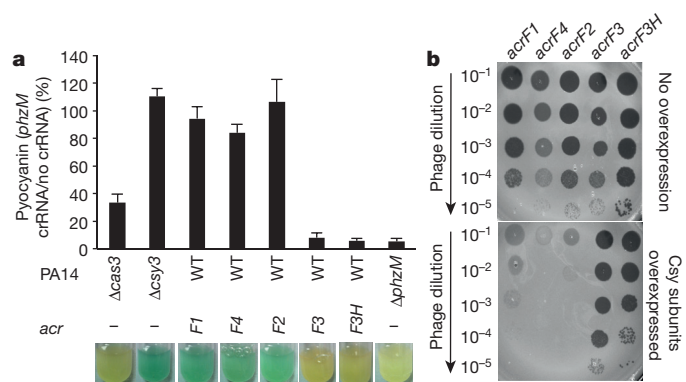


Figure 2 | Anti-CRISPR proteins interact with Cas proteins *in vivo*. **a**, The *phzM* promoter was targeted by a plasmid-encoded crRNA in *P. aeruginosa*. The production of pyocyanin was quantified in different PA14 mutant backgrounds ($\Delta cas3$, $\Delta csy3$ or $\Delta phzM$) or during the expression of the indicated anti-CRISPR from a prophage. The amount of pyocyanin produced in the presence of a plasmid producing the crRNA is shown as a percentage of the same strain with the empty plasmid vector. An average of three independent experiments is shown with error bars representing the standard deviation (s.d.). Representative pictures of cultures are shown. The pyocyanin ratio for the $\Delta phzM$ mutant was derived by comparing it to the value for the $\Delta csy3$ mutant. The prophage expressing *acrF3* also encoded another anti-CRISPR, the functional mechanism of which is not known. To bolster our conclusions pertaining to *acrF3*, we also tested a prophage that expresses an 86% identical homologue of *acrF3*, designated *acrF3H*, and no other anti-CRISPR. WT, wild type. **b**, Lysates of phages expressing the indicated anti-CRISPR proteins were spotted in tenfold serial dilutions on bacterial lawns of wild-type *P. aeruginosa* PA14 (top) or the same strain bearing a plasmid that overexpresses the Csy subunits (bottom). These phages would be targeted by the CRISPR–Cas system in the absence of anti-CRISPR activity.

polymerase chain reaction with reverse transcription (RT-qPCR) experiments showed that these changes in pyocyanin production correlated with reduced transcription of the *phzM* gene (Extended Data Fig. 4). These results demonstrate that the expression of *acrF3* blocks Cas3 activity *in vivo*, causing the Csy complex to function as a transcriptional repressor. Further *in vivo* experiments showed that phages dependent on *acrF1* and *acrF2* for viability⁴ were markedly inhibited by overexpression of the Csy complex subunits (Fig. 2b). The elevated level of Csy proteins probably increases the number of active Csy complexes and/or binds and titrates out anti-CRISPR molecules, resulting in insufficient levels of anti-CRISPR proteins to support robust phage replication. Phages dependent on *acrF3* were not affected under these conditions because this anti-CRISPR protein binds to Cas3, the level of which is unchanged (overexpression of Cas3 inhibited cell growth). Interestingly, Csy subunit overexpression also inhibited a phage expressing *acrF4* (gene 37 from phage JBD26), an anti-CRISPR protein that could not be purified. In addition, expression of this anti-CRISPR in the transcriptional repression assay resulted in a blue-green culture (Fig. 2a). These complementary results imply that AcrF4 binds the Csy complex, which we have experimentally confirmed (Extended Data Fig. 5). We conclude that our *in vivo* experiments are able to distinguish the effects of anti-CRISPR proteins that inactivate the Csy complex from those that inhibit Cas3.

AcrF1 and AcrF2 both prevent DNA binding by the Csy complex, but might achieve this outcome through different mechanisms. The Csy complex assembles with a Csy1–Csy2 heterodimer bound at the 5' end of the crRNA and a Csy4 monomer bound to the 3' end, with six Csy3 subunits arrayed along the backbone of the spacer region in between (Fig. 3a)^{6,8}. By purifying the Csy1–Csy2 heterodimer on its own and mixing it with purified anti-CRISPR proteins, we found that it co-eluted with AcrF2 in SEC experiments, but not with AcrF1 (Fig. 3b and

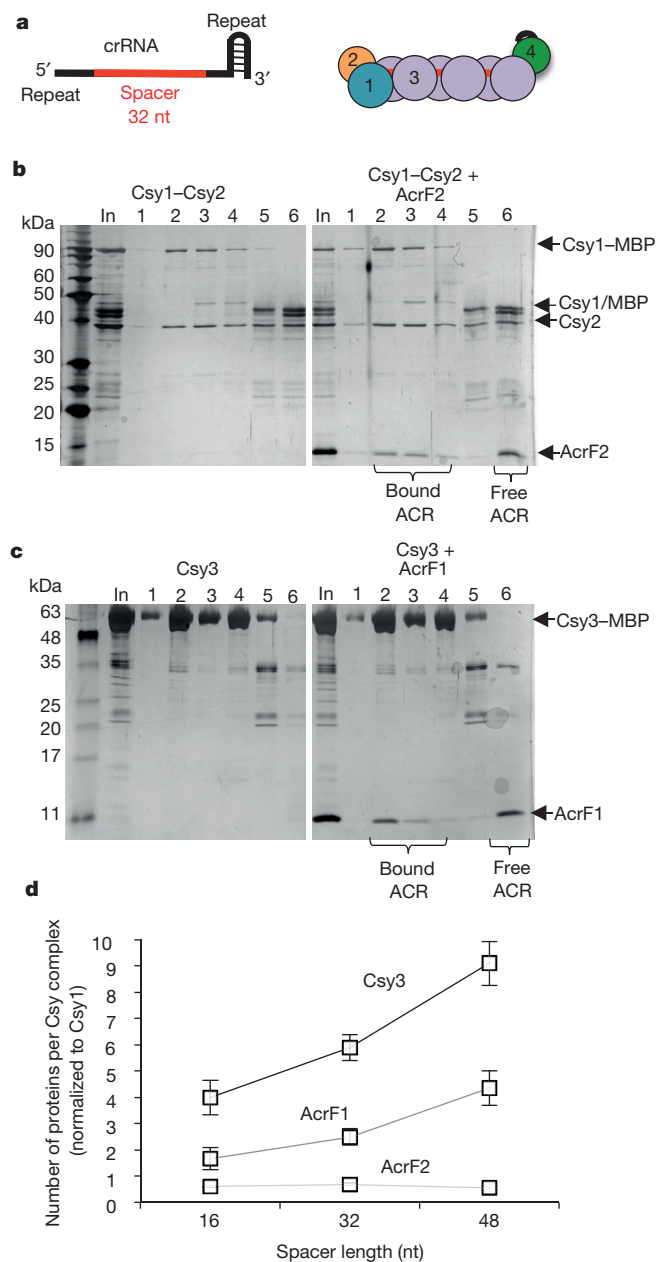


Figure 3 | AcrF1 and AcrF2 bind distinct Csy complex subunits. **a**, A schematic of the crRNA showing the repeat-derived regions of the crRNA (black) and the 32-nucleotide (nt) spacer region (red). The coloured circles represent the Csy1–4 subunits. **b**, **c**, Purified 6×His/MBP-tagged Csy1–Csy2 heterodimer (**b**) or Csy3 (**c**) was fractionated by SEC in the presence (right) or absence (left) of the indicated anti-CRISPR proteins. The SEC fractions were analysed by SDS–PAGE. The ‘In’ lanes show the protein mixture that was loaded onto the SEC column and fractions are numbered. **d**, Purified Csy complexes with 16-, 32-, or 48-nucleotide crRNA spacer regions were bound to AcrF1 or AcrF2 and fractionated by SEC. The stoichiometry of the bound anti-CRISPR proteins was quantified through densitometry of the Coomassie blue stained gels. An average of three independent experiments is shown with error bars representing s.d.

Extended Data Fig. 6a). By contrast, AcrF1, but not AcrF2, bound Csy3 (Fig. 3c and Extended Data Fig. 6b). Csy3 eluted in monomeric and multimeric forms in SEC experiments, with AcrF1 binding predominantly to the multimeric fraction (Fig. 3c). The presence of distinct binding sites for AcrF1 and AcrF2 on the intact Csy complex was corroborated through competition experiments showing that both anti-CRISPR proteins could simultaneously bind the Csy complex and that the presence of one had no effect on the binding ability of the other

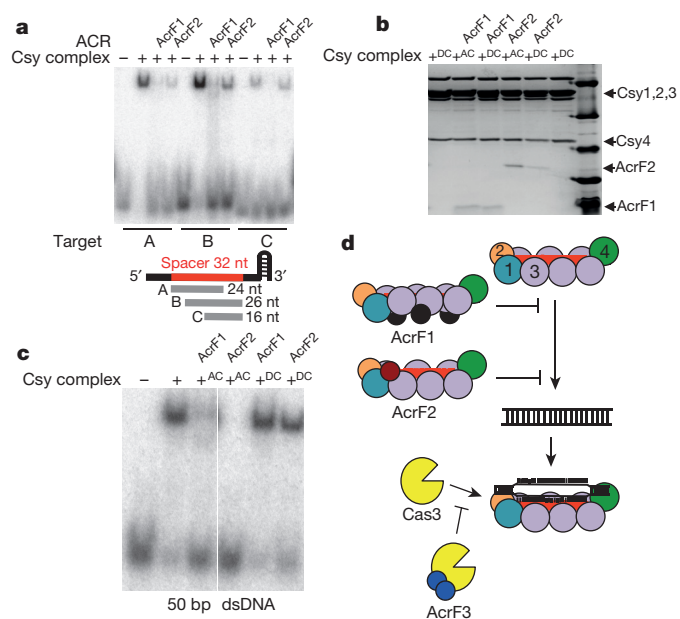


Figure 4 | Two anti-CRISPR proteins inhibit target recognition via unique mechanisms. **a**, EMSA experiments were used to assay binding of the Csy complex to three different ssDNA oligonucleotides (labelled A, B and C) that are complementary to different regions of the crRNA spacer as shown in the schematic (see Extended Data Fig. 9b). Where noted, the Csy complex was pre-incubated with the indicated anti-CRISPR. **b**, **c**, Apo-Csy complex (AC) or DNA-bound Csy complex (DC) was incubated with AcrF1 or AcrF2. **b**, This mixture was fractionated by SEC and fractions were visualized by SDS–PAGE. **c**, An EMSA experiment is shown with binding to dsDNA in the same experimental setup as in **b**. **d**, A model summarizing anti-CRISPR mechanisms. Arrows indicate the steps of the uninhibited CRISPR–Cas interference pathway. Numbers in the Csy complex indicate the Csy subunits. The lines with flat ends indicate the step in the CRISPR–Cas pathway blocked by each anti-CRISPR. The manner in which each anti-CRISPR binds to CRISPR–Cas components is also shown. AcrF1 makes the whole crRNA inaccessible while AcrF2 occludes the 5' end.

(Extended Data Fig. 6c). RNase A treatment of the Csy complex, which resulted in Csy4 dissociation, had no effect on the binding of either anti-CRISPR (Extended Data Fig. 7). Quantification of the co-eluted fractions of AcrF1 or AcrF2 with the Csy complex by protein gel electrophoresis revealed the stoichiometry of AcrF1 to be 2.6 ± 0.3 proteins per Csy complex, while AcrF2 was 0.8 ± 0.1 (Extended Data Fig. 7c). To verify these stoichiometries, we created Csy complexes with shorter (16 nucleotides; Csy₁₆ complex) and longer spacer regions (48 nucleotides; Csy₄₈ complex). The purified Csy₁₆ complex contained fewer molecules of Csy3 (4 ± 0.7) than wild type, and the Csy₄₈ complex contained a proportionally greater number (9 ± 0.8) (Fig. 3d and Extended Data Fig. 8). Concomitant with the altered number of Csy3 molecules in the Csy₁₆ and Csy₄₈ complexes, we observed corresponding changes in the number of AcrF1 molecules bound, with the ratio of Csy3 to AcrF1 remaining constant. These results imply that AcrF1 binds along the full length of the Csy3 ‘spine’ of the complex. Its binding sites are probably at the interaction interfaces of the Csy3 subunits, which would account for the 2:1 Csy3/AcrF1 stoichiometry and for AcrF1 binding to only the multimeric Csy3 fraction (Fig. 3c). In contrast to AcrF1, the number of AcrF2 molecules bound to the altered Csy complexes did not change as the number of Csy3 molecules increased or decreased, consistent with AcrF2 binding to the Csy1–Csy2 heterodimer.

To define further the sites of action of the anti-CRISPR proteins on the Csy complex, we performed DNA-binding assays using ssDNA molecules complementary to subregions of the crRNA spacer. As shown in Fig. 4a, AcrF1 inhibited binding to all the ssDNA molecules tested. By contrast, AcrF2 prevented binding to a 24-nucleotide ssDNA molecule complementary to the 5' end of the crRNA, including the

seed region, but did not inhibit binding to a 16-nucleotide ssDNA complementary to the 3' end of the spacer. Binding to a 26-nucleotide ssDNA binding the 3' end was only partially inhibited. These data suggest that AcrF2 inhibits DNA binding by sterically blocking the 5' end of the crRNA spacer through its interaction with Csy1–Csy2, which is expected to be bound to this region of the crRNA^{17,18}. Addition of AcrF2 to a Csy complex that had been pre-saturated with target DNA resulted in an approximately 60% decrease in the binding level of this anti-CRISPR, suggesting that AcrF2 and DNA compete for an overlapping binding interface (Fig. 4b and Extended Data Fig. 9a). Consistent with this result, addition of AcrF2 to a DNA-bound Csy complex resulted in appreciably decreased DNA binding as detected by EMSA (Fig. 4c). Parallel experiments performed with AcrF1 showed that the binding of AcrF1 to the Csy complex was not affected by prior binding to DNA (Fig. 4b). We conclude that the interaction of AcrF1 with the full length of the spine of the complex formed by multiple Csy3 molecules and the crRNA accounts for its ability to block binding to all dsDNA and ssDNA molecules tested. Furthermore, the ability of AcrF1 and DNA to bind the Csy complex simultaneously suggests an allosteric mechanism for the activity of this anti-CRISPR. Thus, the mechanisms of AcrF1 and AcrF2 are distinct, using different Csy protein-binding partners, stoichiometry and DNA occlusion mechanisms (that is, steric versus allosteric).

We provide the first insight into the mechanisms by which proteins can inhibit a CRISPR–Cas system. The diverse and distinct mechanisms discovered here (Fig. 4d) reflect the deep evolutionary roots of the virus–host arms race. Anti-CRISPR proteins, both known^{4,19} and yet to be discovered, will provide an extensive set of valuable tools both better to understand and to manipulate CRISPR–Cas systems. One example is our finding that AcrF3 converts the CRISPR–Cas system into a gene regulator by blocking Cas3 recruitment. Since CRISPR–Cas systems perform a variety of roles beyond destroying foreign DNA²⁰, many important functions may be fulfilled by proteins that interact with CRISPR–Cas components and thus alter the activity of the system.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 November 2014; accepted 29 July 2015.

Published online 23 September 2015.

1. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
2. Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nature Rev. Microbiol.* **9**, 467–477 (2011).
3. Jore, M. M., Brouns, S. J. J. & van der Oost, J. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. *Cold Spring Harb. Perspect. Biol.* **4**, a003657 (2012).
4. Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493**, 429–432 (2013).
5. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nature Rev. Microbiol.* **12**, 479–492 (2014).

6. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).
7. Wiedenheft, B. *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl Acad. Sci. USA* **108**, 10092–10097 (2011).
8. van Duijn, E. *et al.* Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced short-palindromic-repeats (CRISPR)-associated protein complexes from *Escherichia coli* and *Pseudomonas aeruginosa*. *Mol. Cell. Proteomics* **11**, 1430–1441 (2012).
9. Westra, E. R. *et al.* CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol. Cell* **46**, 595–605 (2012).
10. Huo, Y. *et al.* Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nature Struct. Mol. Biol.* **21**, 771–777 (2014).
11. Rollins, M. F., Schuman, J. T., Paulus, K., Bukhari, H. S. T. & Wiedenheft, B. Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from *Pseudomonas aeruginosa*. *Nucleic Acids Res.* **43**, 2216–2222 (2015).
12. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl Acad. Sci. USA* **108**, 10098–10103 (2011).
13. Wurtzel, O. *et al.* The single-nucleotide resolution transcriptome of *Pseudomonas aeruginosa* grown in body temperature. *PLoS Pathog.* **8**, e1002945 (2012).
14. Luo, M. L., Mullis, A. S., Leenay, R. T. & Beisel, C. L. Repurposing endogenous type I CRISPR–Cas systems for programmable gene repression. *Nucleic Acids Res.* **43**, 674–681 (2015).
15. Rath, D., Amlinger, L., Hoekzema, M., Devulapally, P. R. & Lundgren, M. Efficient programmable gene silencing by Cascade. *Nucleic Acids Res.* **43**, 237–246 (2015).
16. Cady, K. C. & O'Toole, G. A. Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. *J. Bacteriol.* **193**, 3433–3445 (2011).
17. Jackson, R. N. *et al.* Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* **345**, 1473–1479 (2014).
18. Mulepati, S., Héroux, A. & Bailey, S. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484 (2014).
19. Pawluk, A., Bondy-Denomy, J., Cheung, V. H. W., Maxwell, K. L. & Davidson, A. R. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR–Cas system of *Pseudomonas aeruginosa*. *MBio* **5**, e00896 (2014).
20. Westra, E. R., Buckling, A. & Fineran, P. C. CRISPR–Cas systems: beyond adaptive immunity. *Nature Rev. Microbiol.* **12**, 317–326 (2014).

Acknowledgements We thank W. Navarre and E. Westra for reading the manuscript. This work was supported by an Operating Grant to A.R.D. (MOP-130482) and to K.L.M. (MOP-136845), both of which were from the Canadian Institutes of Health Research (CIHR). J.B.-D. was supported by a CIHR Canada Graduate Scholarship Doctoral Award and an Ontario Graduate Scholarship award. Research in the Wiedenheft laboratory is supported by the National Institutes of Health (P20GM103500 and R01GM108888), the National Science Foundation EPSCoR (EPS-110134), the M.J. Murdock Charitable Trust, and the Montana State University Agricultural Experimental Station.

Author Contributions J.B.-D. designed, performed and supervised experiments and wrote the manuscript. B.G., S.S., M.D., and Y.H.-R. performed experiments. M.F.R. and B.W. provided essential reagents and experimental assistance. K.L.M. supervised experiments. A.R.D. designed and supervised experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.R.D. (alan.davidson@utoronto.ca) or K.L.M. (karen.maxwell@utoronto.ca).

METHODS

Protein purification. All proteins were affinity purified using Ni-NTA beads (Qiagen) to isolate recombinant proteins bearing a terminal 6×His tag. Anti-CRISPR proteins were expressed from the p15TV-L vector (NCBI accession number EF456736), which possesses a T7 promoter and an amino-terminal 6×His tag. Constructs expressing Csy1–4 containing a 6×His tag on either Csy3 or Csy4 were co-expressed with a construct producing a crRNA as previously described⁷. Individual Cas proteins (Csy1–Csy2, Csy3, and Cas3) were expressed from pHMGWA (NCBI accession number EU680841), which also has a T7 promoter. The proteins in this vector were tagged with a maltose-binding protein and 6×His.

Cultures of *E. coli* BL21 containing a plasmid expressing a protein of interest were grown to an optical density ($OD_{600\text{ nm}}$) of 0.5 and then induced with 1 mM isopropyl- β -D-thiogalactoside (IPTG) for 3 h at 37 °C (anti-CRISPRs, Csy3) or for 16 h at room temperature (Csy complex, Csy1–Csy2, Cas3). Cells were collected by centrifugation at 5,000g for 10 min and resuspended in a binding buffer (20 mM Tris, pH 7.5, 250 mM NaCl, 5 mM imidazole, 1 mM dithiothreitol (DTT) and 1 mM PMSF). The cells were lysed by sonication and the resulting lysate was centrifuged at 15,000g for 15 min to remove cell debris. The supernatant was mixed with Ni-NTA beads that had been washed in binding buffer (without DTT) five times. Binding to the beads proceeded for 1 h at 4 °C under gentle rotation, at which point the lysate and beads were passed through a column, washed 3–5 times with binding buffer containing 30 mM imidazole and ultimately eluted in buffer containing 250 mM imidazole. Colourimetric Bradford assays were conducted during the procedure to determine the number of washes to perform and elution fractions to collect. Purified protein was dialysed into the binding buffer containing 5 mM imidazole to remove excess imidazole and visualized on Coomassie blue R250 stained SDS–PAGE gels. Cas3 was purified following the same general protocol but in a buffer optimized for this protein (50 mM HEPES, pH 7.5, 500 mM NaCl, 5% glycerol, 1 mM DTT, supplemented with 1 mM PMSF and 150 μ M NiSO_4 in the lysis buffer). Purified Cas3 was concentrated and buffer exchanged in an Amicon Ultra centrifugal filter (Millipore) into a different buffer (20 mM HEPES, pH 7.5, 300 mM KCl, 5% glycerol, 1 mM DTT) for protein interaction assays. Csy1–Csy2 also purified in the same buffer as Cas3 (with NiSO_4 omitted). Purified Csy1–Csy2 was then dialysed into a different buffer (20 mM HEPES, pH 7.5, 250 mM NaCl, 5% glycerol, 1 mM DTT) for protein interaction experiments.

Size-exclusion chromatography. Affinity-purified proteins were fractionated by SEC using a GE Life Sciences Superdex 200 10/300 column. Fractions were collected in 0.5 ml volumes and monitored by optical density at 280 nm. SDS–PAGE gels were stained with silver nitrate or Coomassie blue R250 to identify proteins. In interaction experiments, purified proteins were mixed together before fractionation by SEC and co-eluting proteins were identified by SDS–PAGE. The Csy complex or Csy proteins and an anti-CRISPR protein of interest were generally incubated together for 1 h at 4 °C. This mixture was then applied to the SEC column at room temperature. A fraction of the input (~0.5%) was also kept for SDS–PAGE analysis.

Anti-CRISPR stoichiometry. The purified Csy complex was incubated with ~10-fold molar excess of purified anti-CRISPR proteins. This mixture was fractionated by SEC as described earlier. The Csy complex peak fraction was run on SDS–PAGE gels in twofold serial dilutions. The protein bands were identified with Coomassie blue R250. Image Lab Software (Bio-Rad) was used to quantify band intensities and calculate the relative stoichiometries of the various subunits and anti-CRISPRs, after adjusting for molecular weight and comparing dilutions. Our estimates of the absolute stoichiometries of the Csy subunits is based on the stoichiometry of the Csy complex established in previous publications^{7,8}.

RNase A treatment of the Csy complex. Pancreatic RNase A (73 μ M) was used to treat the Csy complex (4 μ M) for 30 min at 37 °C. After digestion, the treated Csy complex was fractionated by SEC in the absence or presence of an anti-CRISPR protein. Fractions from SEC were analysed on Coomassie stained SDS–PAGE gels to visualize proteins and SYBR Gold stained TBE–Urea gels to visualize nucleic acid.

Isothermal titration calorimetry. Purified Csy complex was added to the isothermal titration calorimetry (ITC) chamber at a concentration of 7.5 μ M. The DNA ligand (8-nucleotide ssDNA) was placed in the injection syringe at a concentration of 75 μ M. After a null injection of 0.3 μ l of titrant, 3 μ l of titrant were injected 13 times, with 120 s intervals between the injections to establish a baseline. The DNA titrant and Csy complex were in the same buffer (20 mM Tris, pH 7.5, 250 mM NaCl, 5 mM imidazole) and the experiment was temperature controlled at 25 °C. To assess the role of AcrF1 in interfering with the interaction between the Csy complex and a DNA target, the Csy complex was first incubated with a ~10-fold molar excess of anti-CRISPR proteins for 1 h at 4 °C. This mixture was then applied to the chamber, the temperature equilibrated to 25 °C and the DNA titration performed.

Electrophoretic mobility shift assay. A 50-nucleotide ssDNA molecule was synthesized (Eurofins Genomics) that contains 32 nucleotides of complementarity to the crRNA in the purified Csy complex. The DNA (200 nM) was phosphorylated in a T4 polynucleotide kinase reaction with [γ -³²P]ATP. The reaction was stopped with 12 mM EDTA and GE MicroSpin G-25 columns were used to remove remaining radiolabelled nucleotides. To generate dsDNA, the labelled strand was heated to 98 °C in the presence of a twofold excess of an unlabelled complementary strand and allowed to return slowly to room temperature. Csy complex–DNA-binding reactions were conducted in a binding buffer (10 mM HEPES, pH 7.5, 1 mM MgCl_2 , 20 mM KCl, 1 mM TCEP, bromophenol blue and 6% glycerol) at 37 °C for 15 min. The concentration of the Csy complex used in EMSA experiments varied, depending on the oligonucleotide target being used. For 50 bp dsDNA EMSA reactions, 100 nM of the Csy complex was routinely used in reactions, with <1 nM labelled DNA. Anti-CRISPR proteins were used at a tenfold molar excess compared to the Csy complex and allowed to incubate with Apo-Csy complex or DNA-bound Csy complex for 1 h. After the appropriate incubation, the reactions were resolved on native 6% polyacrylamide TBE gels. Gels were wrapped in Saran wrap and visualized with a phosphoscreen and Typhoon imager. Optimal exposures were ~2–3 h.

For EMSA experiments involving Cas3, the Csy complex and target DNA were prebound as described above. 6×His-tagged Cas3 was purified by Ni-NTA chromatography (6×His) followed by SEC, concentrated, transferred into the EMSA reaction buffer, flash frozen in small volumes (50 μ l) and stored at –70 °C. Cas3 was added to the EMSA reaction at a final concentration of 400 nM and incubated for 30 min at 37 °C. ATP was added at a final concentration of 2 mM and all reactions with Cas3 also contained 100 μ M CoCl_2 .

Pyocyanin repression. A crRNA was designed to target the promoter region of *phzM*, a gene required for the biosynthesis of the blue–green pigment pyocyanin. Two complementary oligonucleotides were synthesized containing two 28 bp PA14 CRISPR repeat sequences, flanking a 32 bp sequence with perfect complementarity to the –35/–10 region of the *phzM* promoter (position 813576–813607 in the PA14 genome). The spacer was designed to produce a crRNA that would bind to the non-template strand, in a position where the protospacer adjacent motif (GG) is present. The oligonucleotides were annealed and cloned into an arabinose inducible *P. aeruginosa* expression vector, pHERD30T. This construct was then used to transform PA14 strains possessing single *cas* gene knockouts or wild-type PA14 possessing prophages expressing various anti-CRISPRs. Individual transformants were grown overnight (~20 h) in 2 ml of King's A media in 50 μ g ml^{–1} gentamicin and 0.025% arabinose, to induce expression of the crRNA. Pyocyanin was extracted with an equal volume of chloroform, and then mixed with 1 ml of 0.2 M HCl, producing a pink–red colour proportional to the amount of pyocyanin, which was quantitated by measuring absorbance at 520 nm. Anti-CRISPR proteins were expressed from the following prophages: JBD30 (AcrF1), D3112 (AcrF2), JBD26 (AcrF4), JBD5 (AcrF3 and AcrF5), and JBD88a (AcrF3H). Since phage JBD5 contains two type I-F anti-CRISPR proteins, phage JBD88a (possessing a homologue of AcrF3 with 86% protein sequence identity) was also used.

Competition experiments. To determine whether the two anti-CRISPR proteins that bind to the Csy complex compete with each other for the same binding site, the first anti-CRISPR was added for 1 h at 4 °C and then the second for the same amount of time. This entire mixture was then fractionated by SEC.

To determine whether DNA and anti-CRISPR proteins compete for the same binding site, the purified Csy complex (4.5 μ M) was mixed with a 50 bp dsDNA target (10 μ M) and incubated for 15 min at 37 °C in the same buffer in which the proteins were purified (20 mM Tris, pH 7.5, 250 mM NaCl, 5 mM imidazole). This DNA-bound Csy complex was then mixed with a tenfold molar excess of AcrF1, AcrF2, or an equivalent volume of buffer and incubated for 1 h at 4 °C. This mixture was fractionated by SEC. The fraction containing the Csy complex was analysed on Coomassie blue stained SDS–PAGE gels or SYBR Gold stained TBE–Urea gels.

Plaque assays with Csy subunit overexpression. To assess the consequence of Csy protein overexpression on phages possessing distinct anti-CRISPR proteins *in vivo*, a pHERD30T derived plasmid expressing the *csy1*, *csy2*, *csy3* and *csy4* genes was used to transform *P. aeruginosa* strain PA14. Phage lysates were spotted in tenfold serial dilutions onto a lawn of PA14 containing empty vector, or the plasmid expressing the *csy* genes. Phages JBD30, JBD26, D3112 and JBD88a all have protospacers that display 100% identity to spacers 17 and 20 in the PA14 CRISPR2 locus⁴. JBD5 has a protospacer matching CRISPR2 spacer 1 that has been shown to be targeted^{4,21}.

RT–qPCR. RT–qPCR reactions were conducted as described previously⁴. Briefly, total RNA was extracted and DNase treated. One nanogram of total RNA was subjected to a reverse transcription reaction and qPCR, using primers specific to *phzM* or a control, *rpsL*. The efficient removal of DNA from the RNA

preparation was confirmed by including controls for each sample without reverse transcriptase added.

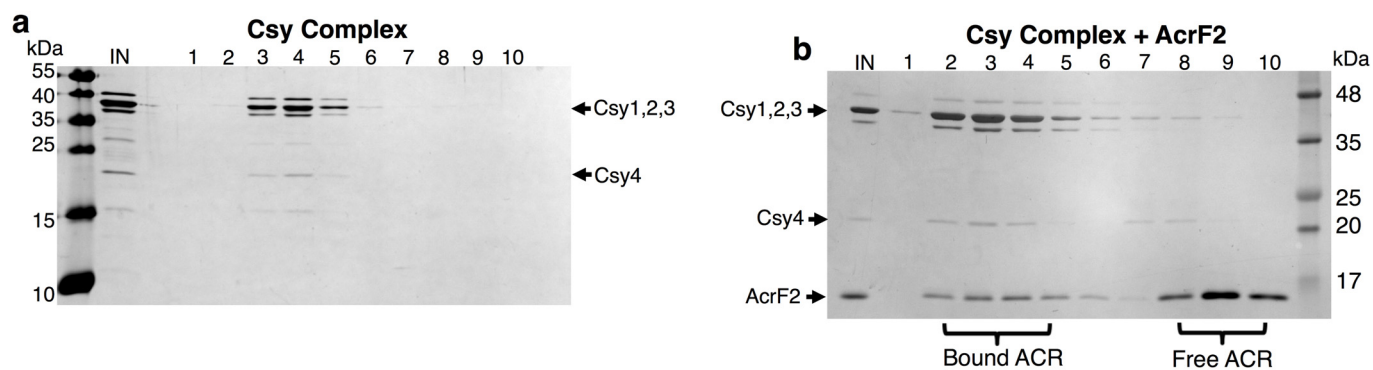
AcrF2 misannotation. The D3112 phage genome has an annotated open reading frame identified as gene 30, which is a predicted 90 amino acid protein (NCBI accession number NC_005178). This version of the gene was previously identified as an anti-CRISPR, although overexpression from a plasmid was required for activity⁴. A nucleotide alignment of the anti-CRISPR region of many phages revealed that all phage anti-CRISPR operons possess a start codon (ATG) at the same position for the first anti-CRISPR gene, except phage D3112. Phages D3112 and MP29 (which has a D3112 gene 30 homologue), had the start position annotated downstream of this commonly used ATG, at a second ATG, in frame with the first, resulting in a putative truncation of six amino acid residues. Re-cloning of the gene to include these six residues resulted in a construct that had full anti-CRISPR activity in the absence of overexpression. Thus, this 96-residue protein (sequence shown later, with new residues in bold) is the version that was used in all downstream experiments presented here and in affinity purification, after addition of the appropriate tag. All other anti-CRISPR protein sequences are as reported in ref. 4. AcrF2: **MTKTAQMIAQHKDTVAACEAAEAIAIAKDQVWDGEGYT**

KYTFDDNSVLIQSGTTQYAMDADDADSIKGYADWLDDEARSAEASEIER
LLESVEEE.

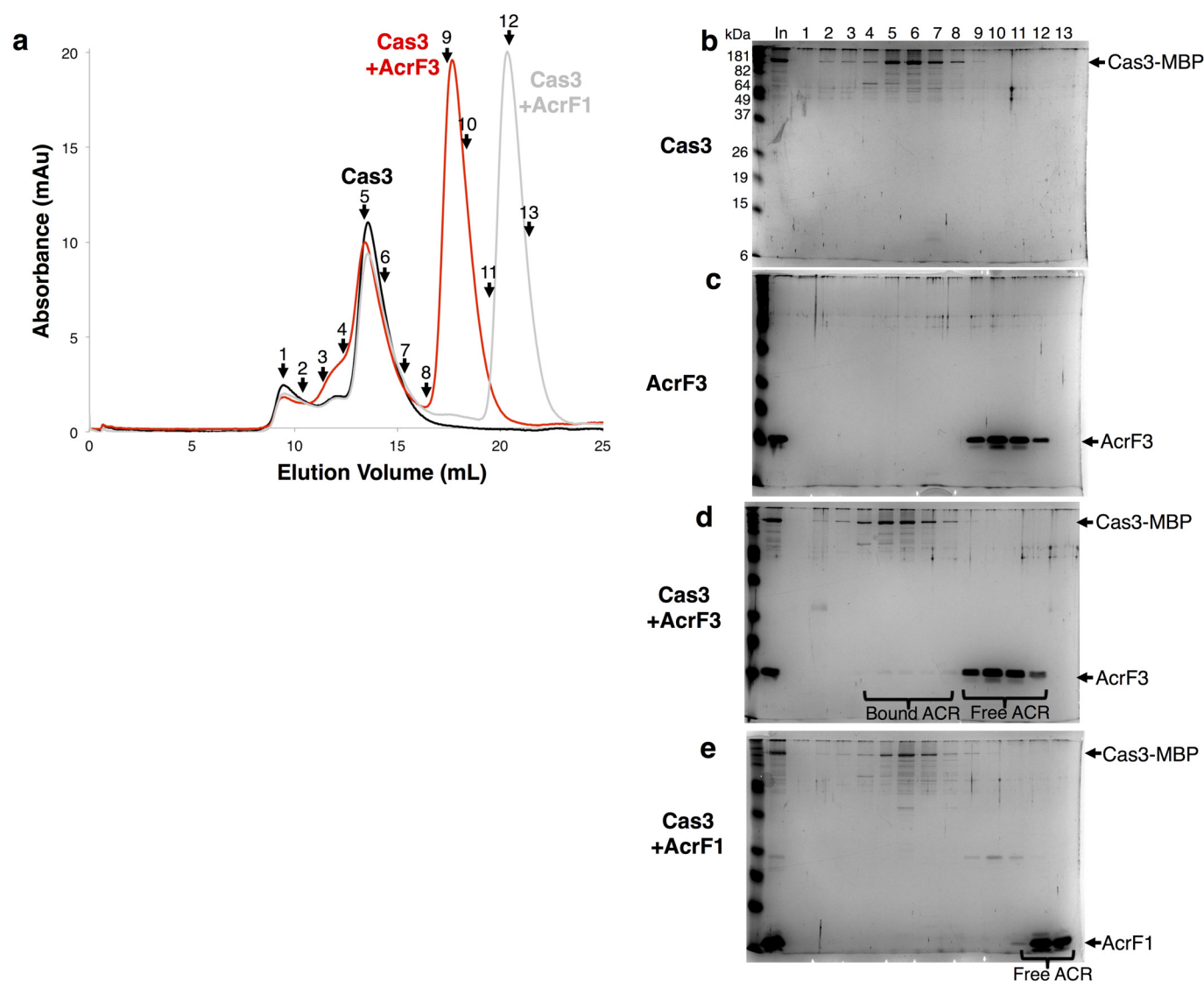
Statistics, reagents and data deposition. To assess interactions between anti-CRISPR proteins and the Csy complex or purified Cas proteins, mixed components were fractionated by SEC. Each result shown in the manuscript was obtained on at least two independent occasions. ITC, EMSA and plaque assays were all replicated at least three times. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

The sequences of the anti-CRISPR proteins are present in ref. 4, with full genomes for phages JBD30, D3112, JBD5, JBD26 and JBD88a available on NCBI (accession numbers: NC_020198, NC_005178, NC_020202, JN811560 and NC_020200, respectively).

21. Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O'Toole, G. A. The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *J. Bacteriol.* **194**, 5728–5738 (2012).

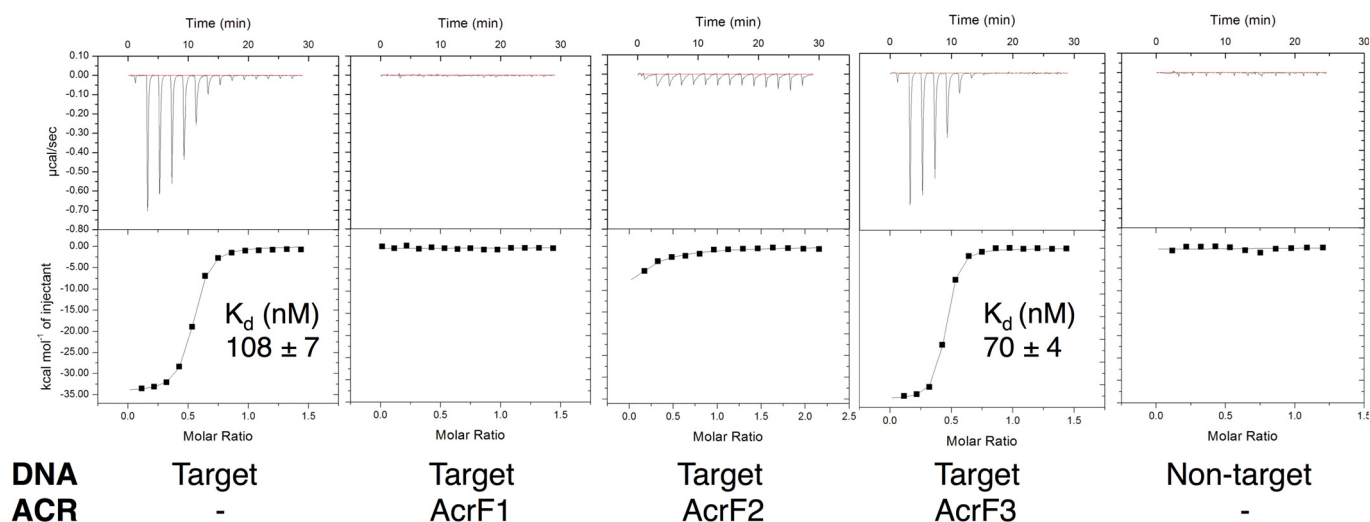


Extended Data Figure 1 | AcrF2 interacts with the Csy complex. a, b, Purified Csy complex was fractionated by SEC alone (a) or in the presence of AcrF2 (b). Fractions were analysed on a silver nitrate stained SDS-PAGE gel. The input (IN) and fractions are shown.



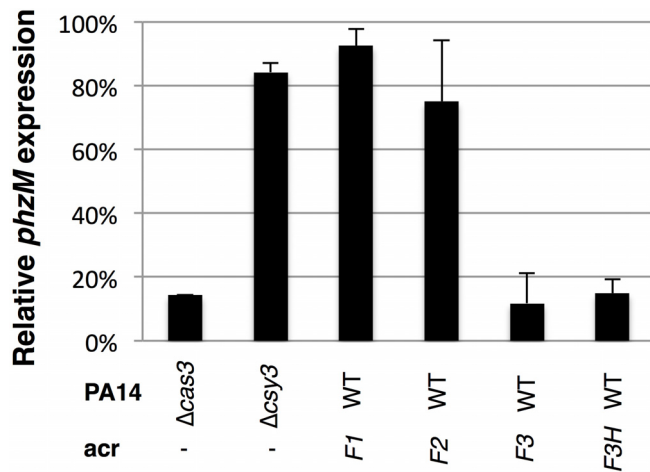
Extended Data Figure 2 | AcrF3, not AcrF1, interacts with Cas3. **a**, Cas3 was fractionated by SEC alone or in the presence of AcrF3 or AcrF1. Overlays of plots of elution volume versus optical density at 280 nm of the column eluates are shown. The numbers represent the fractions that were selected for analysis. **b–e**, Silver nitrate stained SDS–PAGE gels are shown from SEC experiments with Cas3 (**b**), AcrF3 (**c**), Cas3 with AcrF3 (**d**) or Cas3 with AcrF1 (**e**).

The sample that was loaded onto the SEC column is shown as input (In) and fractions from the same elution positions are indicated numerically. AcrF3 is seen eluting in fractions 4–8 only in the presence of Cas3. There is also a visible shift in the Cas3 elution profile in the presence of AcrF3 but not AcrF1 (fractions 3–5).

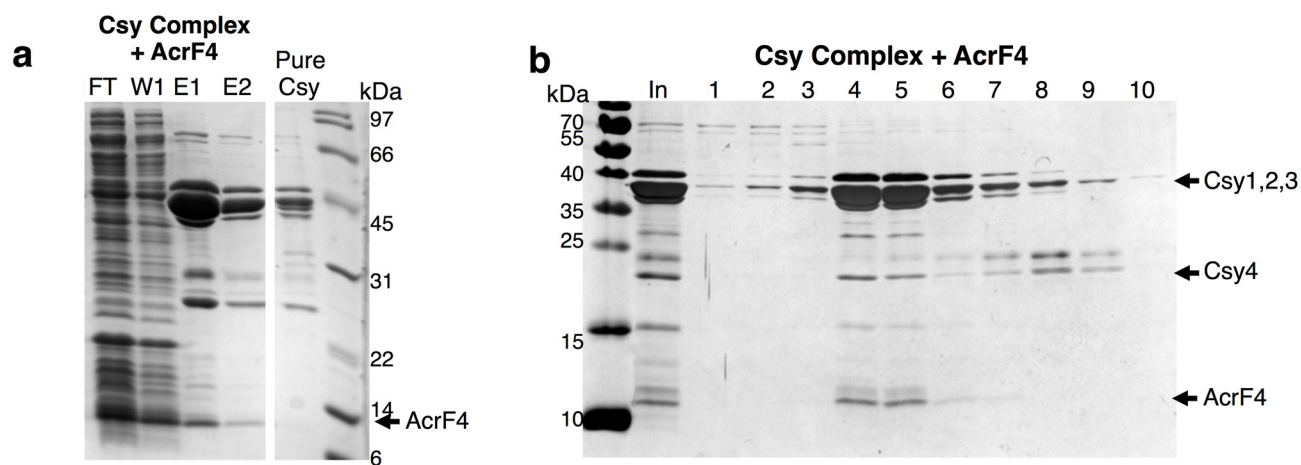


Extended Data Figure 3 | AcrF1 and AcrF2 prevent target recognition by the Csy complex. Isothermal titration calorimetry (ITC) assays showing the Csy complex binding to an 8-nucleotide ssDNA target that comprises the seed region. No binding is observed in the presence of AcrF1, AcrF2 or with a non-target (the reverse complement sequence of the target) ssDNA substrate.

A representative run is shown for each condition with the dissociation constant (K_d) value and error of fit from that particular run. Over multiple runs ($n = 6$) with the Csy complex binding to the ssDNA ligand, the average K_d value was $90 \text{ nM} \pm 37$.

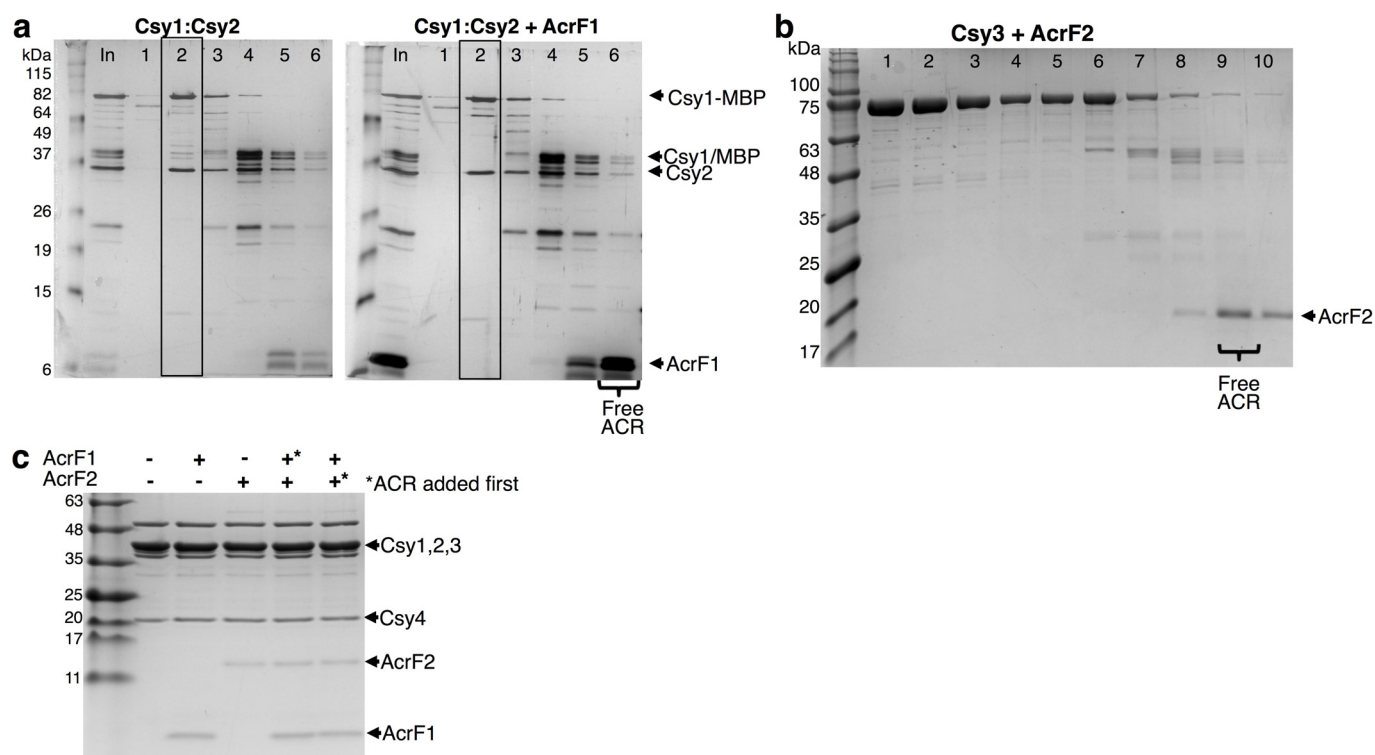


Extended Data Figure 4 | Expression of *phzM* is repressed by the Csy complex. The Csy complex was targeted to the promoter of the gene *phzM*, and repression efficiency was assayed by RT-qPCR (see Methods). The per cent repression of *phzM* in the indicated strains expressing a *phzM*-targeting crRNA relative to wild-type (WT) PA14 with an empty plasmid is shown. All values were normalized to *rpsL*, a gene encoding a ribosomal protein. Means \pm s.d. are shown.



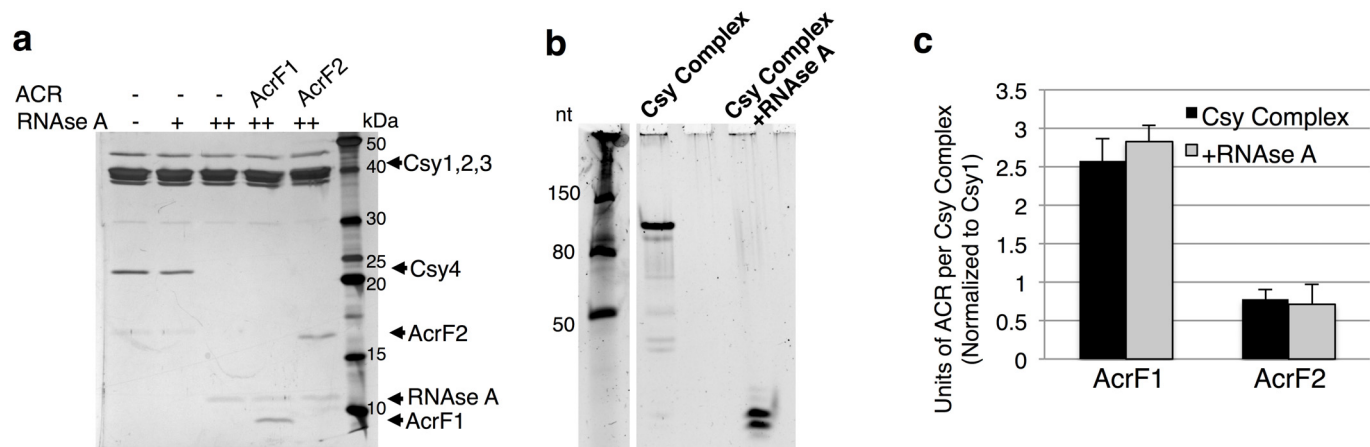
Extended Data Figure 5 | AcrF4 interacts with the Csy complex. Untagged AcrF4 was expressed in *E. coli* BL21 cells and a crude lysate of these cells was mixed with the Csy complex bound to Ni-NTA beads via a 6×His tag on Csy3. **a**, The flow through (FT), wash 1 (W1), and two elution fractions (E1, E2) from the Ni-NTA column are shown, as well as a comparison to pure Csy

complex. **b**, The Ni-NTA elution fractions were fractionated by SEC, demonstrating a stable interaction between the Csy complex and AcrF4. The input (In) lane shows the sample that was loaded on the SEC column and numbered fractions are analysed on SDS-PAGE gels.



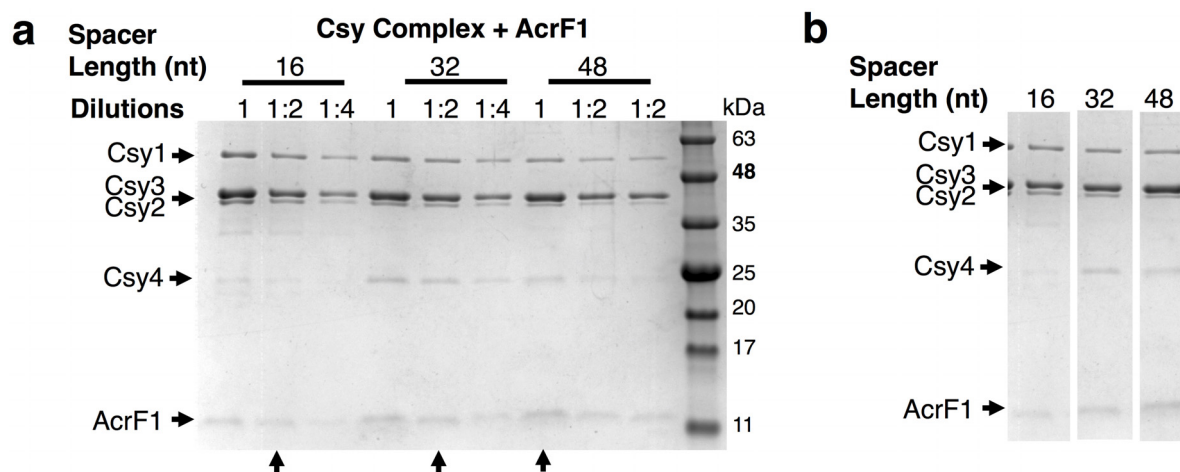
Extended Data Figure 6 | AcrF1 and AcrF2 bind the Csy complex at distinct locations. **a**, Purified Csy1–Csy2 heterodimer with an MBP and 6×His tag fused to Csy1 was fractionated by SEC in the presence or absence of AcrF1 (boxes indicate the Csy1–Csy2 peak). **b**, Purified MBP/6×His-tagged Csy3 was fractionated in the presence or absence of AcrF2. These are complementary experiments to those seen in Fig. 3b and c, respectively. Input (In) and selected fractions are shown on an SDS–PAGE gels. **c**, AcrF1 and AcrF2 were incubated

with the Csy complex singly or in combination. Asterisks designate which anti-CRISPR was added first to the reactions containing both anti-CRISPR proteins. The addition order did not affect the result since there is no competition for binding sites between these two anti-CRISPR proteins. After incubation, each mixture was fractionated by SEC and the peak Csy complex fraction is shown on an SDS–PAGE gel. In each experiment the anti-CRISPR proteins are in excess relative to the Csy complex.



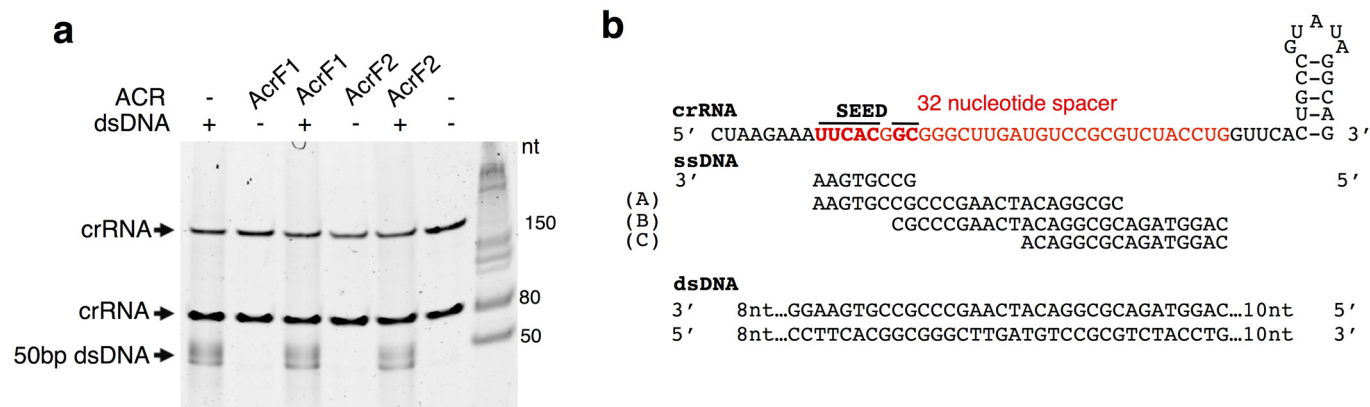
Extended Data Figure 7 | AcrF1 and AcrF2 interact with an RNase-A-treated Csy complex. **a**, The Csy complex was treated with a low concentration (600 nM, +) of RNase A or a high concentration of RNase A (70 μ M, ++). This mixture was fractionated by SEC, revealing Csy4 dissociation at the higher RNase A concentration. Pre-treatment of the Csy complex with RNase A, with the subsequent addition of AcrF1 or AcrF2 followed by SEC fractionation was then conducted. Peak Csy complex fractions are shown on an SDS-PAGE

gel. **b**, A TBE-urea denaturing gel is shown, stained with SYBR gold, showing the native crRNA in the Csy complex and the protected fragments remaining after 70 μ M RNase A treatment. **c**, Quantification of Coomassie blue stained gels from three independent preparations of the respective proteins is shown. Anti-CRISPR proteins bound with unaltered stoichiometry to RNase-A-pre-treated Csy complexes. Error bars represent s.d.



Extended Data Figure 8 | Twofold dilutions used to quantify anti-CRISPR binding stoichiometry. Csy complexes with crRNA molecules possessing spacers of differing lengths (16, 32, or 48 nucleotides) were purified and fractionated by SEC in the presence of AcrF1. A representative Coomassie blue stained SDS-PAGE gel is shown, with twofold dilutions of the peak fraction

containing the Csy complex and co-eluting AcrF1. Arrows on the bottom of the gel indicate comparable dilutions based on the levels of Csy1. Note the increasing abundance of Csy3 and AcrF1. **b**, Lanes with arrows from the gel in **a** are shown next to each other for comparison.



Extended Data Figure 9 | dsDNA binds to the Csy complex after SEC fractionation. **a**, The same samples from Fig. 4a were run on a denaturing TBE-urea gel, stained with SYBR gold, to reveal the crRNA (two species are apparent), and the Csy-complex-bound 50 bp dsDNA. In these experiments, DNA was prebound to the Csy complex, and AcrF1 or AcrF2 were subsequently added to the DNA-saturated Csy complex. This mixture was then fractionated by SEC and the Csy-complex-containing peak fractions were

analysed. **b**, A schematic showing the crRNA sequence with repeat-derived regions shown in black and the variable 32-nucleotide spacer region in red. The seed-interacting region that is critical for target recognition (nucleotides 1–5, 7, 8) is in bold. DNA oligonucleotides used in this study are shown, with labels 'A', 'B' and 'C' corresponding to the targets shown in Fig. 4c. The 8-nucleotide ssDNA substrate was used in ITC experiments (Extended Data Fig. 3), and the 50 bp dsDNA in EMSAs (Figs 1d and 4b).

In situ structural analysis of the human nuclear pore complex

Alexander von Appen^{1*}, Jan Kosinski^{1*}, Lenore Sparks^{1*}, Alessandro Ori¹, Amanda L. DiGuilio², Benjamin Vollmer^{3†}, Marie-Therese Mackmull¹, Niccolo Banterle¹, Luca Parca¹, Panagiotis Kastiris¹, Katarzyna Buczak¹, Shyamal Mosalaganti¹, Wim Hagen¹, Amparo Andres-Pons¹, Edward A. Lemke¹, Peer Bork¹, Wolfram Antonin³, Joseph S. Glavy², Khanh Huy Bui^{1,4} & Martin Beck¹

Nuclear pore complexes are fundamental components of all eukaryotic cells that mediate nucleocytoplasmic exchange. Determining their 110-megadalton structure imposes a formidable challenge and requires *in situ* structural biology approaches. Of approximately 30 nucleoporins (Nups), 15 are structured and form the Y and inner-ring complexes. These two major scaffolding modules assemble in multiple copies into an eight-fold rotationally symmetric structure that fuses the inner and outer nuclear membranes to form a central channel of ~60 nm in diameter¹. The scaffold is decorated with transport-channel Nups that often contain phenylalanine-repeat sequences and mediate the interaction with cargo complexes. Although the architectural arrangement of parts of the Y complex has been elucidated, it is unclear how exactly it oligomerizes *in situ*. Here we combine cryo-electron tomography with mass spectrometry, biochemical analysis, perturbation experiments and structural modelling to generate, to our knowledge, the most comprehensive architectural model of the human nuclear pore complex to date. Our data suggest previously unknown protein interfaces across Y complexes and to inner-ring complex members. We show that the transport-channel Nup358 (also known as Ranbp2) has a previously unanticipated role in Y-complex oligomerization. Our findings blur the established boundaries between scaffold and transport-channel Nups. We conclude that, similar to coated vesicles, several copies of the same structural building block—although compositionally identical—engage in different local sets of interactions and conformations.

Throughout eukaryotes, the nuclear pore complex (NPC) has a three-ringed, sandwiched architecture¹. The inner ring complex (IR) has five members in mammals, namely Nups 205, 188, 155, 93 and 35 (ref. 2). The cytoplasmic and nuclear rings (CR and NR, respectively) are composed of multiple copies of the ten-membered Y complex. X-ray structures of almost all Y complex domain folds have been solved³ and their localization within the overall Y shape has been determined *in vitro*^{4–6}. Nup160, Nup37 and Elys localize to the large arm of the Y, and Nup85, Nup43 and Seh1 to the small arm. Nup96 and Sec13 comprise the stem base and join the two arms in a central hub element, collectively known as the vertex. Nup107 and Nup133 comprise the stem tip and are highly flexible in isolation (Fig. 1b)^{4,5}. Scaffold Nups contain β -propeller and α -solenoid domains¹. Although the NPC scaffold is symmetric across the nuclear envelope plane, a different set of transport-channel Nups asymmetrically binds to the CR and NR. Two subcomplexes exclusively localize to the CR, the Nup214–Nup88–Nup62 complex⁷ that also contains Rae1 and Nup98, as well as the Nup358–RanGAP1*SUMO1–Ubc9 complex⁸. Elys, Tpr, Nup50 and Nup153 are well-established members of the

NR¹. In contrast, the Nup62–Nup58–Nup54 complex that is anchored to the scaffold via Nup93¹, appears symmetrically distributed.

We previously proposed that a total of 32 Y complexes form two concentric, reticulated rings in both the CR and NR^{4,9}. The basic structural element is a dimer of two slightly shifted, tandem complexes that are referred to as the outer and inner Y complex with respect to their distances to the nucleocytoplasmic axis. A recent crystallographic analysis of the entire vertex was overall consistent¹⁰. From these findings arose a number of questions about how exactly this intriguing

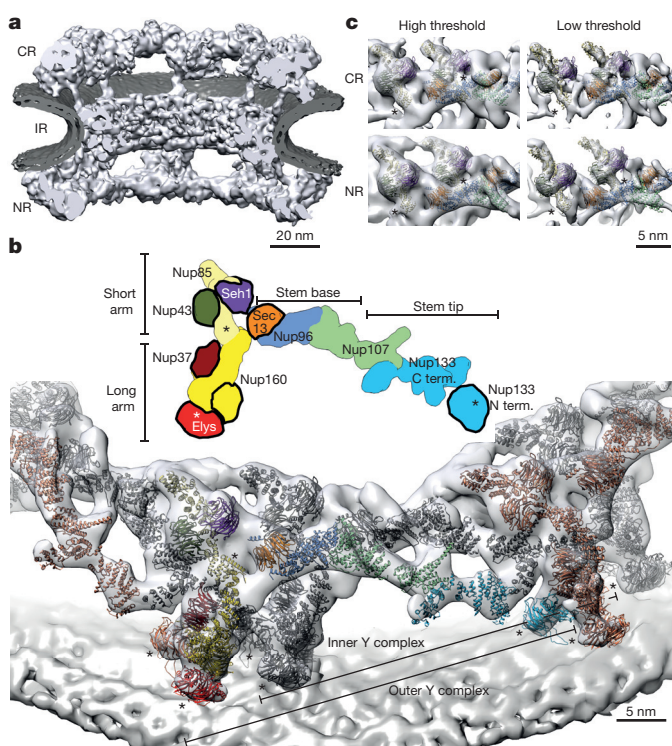


Figure 1 | Tomographic map of the human NPC. **a**, The structure is shown cut in half (membranes in dark grey). **b**, Segment of the NR with the two staggered inner (grey) and outer Y complexes with their anterior and posterior counterparts (colored grey and orange). Colour code corresponds to the scheme on top (β -propellers highlighted with black strokes). Asterisks mark structures that can be unambiguously positioned but have some uncertainty in their orientation. **c**, The small arm region of the inner and outer vertices is shown with two different isosurface thresholds. The fits of Nup85–Seh1, Nup43 and Sec13–Nup96–Nup107 are shown superimposed.

¹European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany. ²Department of Chemistry, Chemical Biology and Biomedical Engineering, Stevens Institute of Technology, 507 River St., Hoboken, New Jersey 07030, USA. ³Friedrich Miescher Laboratory of the Max Planck Society, Spemannstrasse 39, 72076 Tübingen, Germany. ⁴Department of Anatomy and Cell Biology, McGill University, Montreal, Quebec H3A 0C7, Canada. [†]Present address: Oxford Particle Imaging Centre, Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.

*These authors contributed equally to this work.

architectural arrangement is held together. Specifically, how are two concentric rings of different diameter formed by the same Y-complex building blocks? How do they interact with each other? Are those interactions identical within the NR and CR? Are they sufficient for the stability of the two concentric rings or are additional Nups required for oligomerization? Here, we address these questions, using an integrated *in situ* structural biology approach.

Using direct electron detection, we obtained a tomographic map of the NPC resolved overall to 23 Å with some local deviations in resolution (Fig. 1a and Extended Data Figs 1 and 2). In each of the four vertex positions per asymmetric unit, the crystal structures or homology models are seamlessly fitted (Fig. 1b), and certain structural features, for example, individual β -propellers, are clearly evident (Fig. 1c, Extended Data Figs 2b and 3, Extended Data Table 1, Supplementary Table 1 and Supplementary Video 1). In addition to previous analysis^{4,10} the structures of Nup37, Nup43 and the Nup107 C-terminal domain in complex with the Nup133 C-terminal domain can be unambiguously assigned outgoing from the vertex proteins. These assignments are critical in understanding how Y complexes interact with each other *in situ* (see below). Although electron optical density for the β -propellers of Elys and Nup133 N-terminal domain is clearly seen in proximity of their known interfaces to Nup160^{11,12}, their exact orientation remains partially ambiguous because they might rotate around the β -propeller axis or flip perpendicularly to it (Supplementary Table 1). These interfaces—although biochemically defined—still await high-resolution structural analysis.

A number of minor conformational differences between the inner and outer Y complexes are apparent *in situ* (Extended Data Fig. 4), which are consistent with the hinges observed within the stem *in vitro*^{4,13}. A major conformational difference is apparent for the N-terminal β -propeller of Nup133 that is important for forming the head-to-tail contact with Nup160 (ref. 12). Although both the inner and the outer copy are identically positioned with respect to Nup160, their position with respect to the corresponding Nup133 C-terminal domain is very different (Extended Data Fig. 4d), suggesting a critical function of the hinges for scaffold formation. Our structural model suggests five protein interfaces that contribute to Y-complex oligomerization *in situ* (Extended Data Fig. 4d), only two of which were previously described^{4,10,12,14}. Interestingly, several of those across the inner and outer Y complex appear to be specific to higher eukaryotes. The inner Nup133 (residues 519–907) interfaces with the outer Nup37 and a region of Nup160 of unknown structure (residues 181–195). The outer Nup43 interfaces with the inner finger¹⁵ region of Nup107 (Extended Data Figs 2b and 4e). The latter four are absent in *Saccharomyces cerevisiae*. Furthermore, phosphorylation sites that might control NPC disassembly at the onset of mitosis in vertebrates^{16,17} are enriched at such interfaces (Extended Data Fig. 4f). These findings support the previously proposed hypothesis¹⁸ that NPC scaffold organization might be variable across species.

The stem bases of inner and outer Y complexes are differently connected in the CR as compared to NR (Fig. 2a, b and Extended Data Fig. 5). In the CR, this density possibly could account for either of the Nup214 or 358 complexes. Gene silencing of the Nup214 however does not cause alterations of Y complexes⁴. To test if Nup358 complex contributes to this density, we obtained a tomographic map after removal of Nup358 complex using gene silencing (Extended Data Fig. 5). This structure displayed a striking phenotype. The outer Y complex was missing in the CR (Fig. 2c, d) but not in the NR (Extended Data Fig. 5b). This finding was confirmed using classification of sub-tomograms and stoichiometric measurements (Fig. 2e and Extended Data Fig. 5e, f). We conclude that Nup358, which is not yet classified as a scaffold Nup, plays an unanticipated role in the maintenance of the quaternary structure of the CR. Intriguingly, variations in expression levels of Nup358 are highly relevant to human health (for example, see ref. 19). Which Nup fulfils a similar role in the NR remains to be investigated. Interestingly, a recent

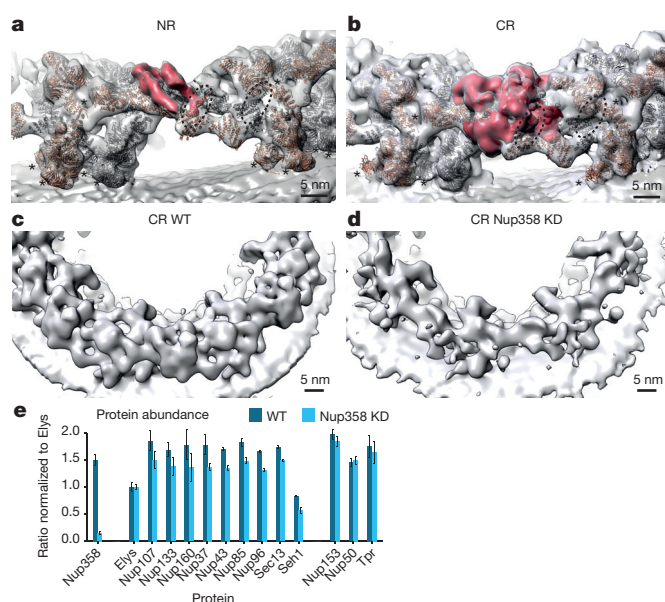


Figure 2 | Nup358 complex stabilizes the CR. **a, b**, The connecting density (red) between the inner (grey) and outer (orange) Y complexes differs in the NR (**a**) as compared to the CR (**b**). While the outer finger domain of Nup107 (dotted lines) is engaged with the connecting density, its inner counterpart is engaged with Nup43. Asterisks mark structures that can be unambiguously positioned but have some uncertainty in their orientation. **c, d**, Removal of Nup358 complex causes loss of the outer Y complex in the CR but not NR (Extended Data Fig. 5). Tomographic structures are shown for the wild-type (WT, **c**) and knockdown condition (KD, **d**). **e**, Stoichiometric measurements show that in contrast to nuclear-oriented Nups, Y-complex members are reduced by ~20% upon Nup358 knockdown (combined *P* value excluding Elys: 1.31×10^{-6} ; one-sided Welch *t*-test combined using Fisher's method; error bars indicate median absolute deviation across three biological replicates and multiple independent peptides).

proximity labelling study suggested that Tpr binds to the respective region of the Y complex²⁰.

Additional density that cannot be attributed to members of the Y complex but is consistent between the NR and CR was apparent in two specific regions. First, a peculiar question-mark-shaped density is evident in proximity to the small arm of the outer vertex in both the NR and CR (Fig. 3a and Extended Data Fig. 6a), which would be consistent with the established shape of certain IR scaffold Nups²¹. It is positioned such that it forms multiple contacts to both vertices and connects them to the inner stem of the anterior asymmetric unit, implying a critical role in ring formation. Second, a rod-shaped density that connects the inner vertex with the IR complex is symmetric across the nuclear envelope plane (Fig. 3b). To systematically explore structural similarity between the tomographic map and the remaining IR scaffold Nups, we performed a systematic fitting of the available structures of Nups 155 and 205 and 188; the latter two are paralogous and cannot be distinguished at the given resolution. This analysis was consistent with a positioning of Nups 205/188 into the question-mark-shaped density in the CR and NR as well as Nup155 in the rod-shaped connector, but also identified several hits within the IR, as expected (Fig. 3e and Extended Data Fig. 7). We biochemically confirmed weak interactions of Nups 93, 155 and 205 with the Y complex (in interphasic cells; Extended Data Fig. 8), and a strong interaction of Nup188 with the Nup214 complex (in nocodazole-arrested, that is, mitotic cells; Fig. 3c and Extended Data Fig. 8). The Nup214 complex is known to bind to the inner vertex region of the CR⁴, close to the aforementioned question-mark shape. In the case of Nup155, the orientation of the fitted structure suggests that a specific loop of its β -propeller domain dips into the lipid bilayer (Fig. 3b). Liposome binding experiments with wild-type *Xenopus* Nup155 and

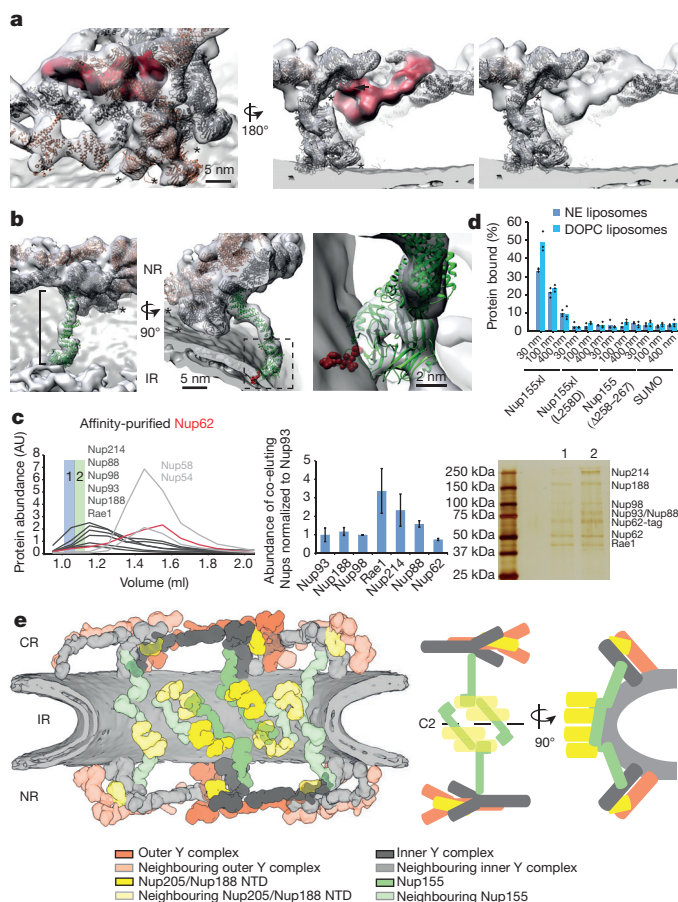


Figure 3 | Scaffold architecture of the human NPC. **a**, Question-mark-shaped density (red) in the vertex region of the NR (arrowhead indicates Nup43). **b**, Fit of Nup155 (green) into the rod-shaped density connecting the outer rings with the IR. The membrane-binding motif (red) dips into the outer lipid bilayer. Asterisks mark structures that can be unambiguously positioned but have some uncertainty in their orientation. **c**, Nup188 and Nup93 co-purify with the Nup214 complex. Nup62 was affinity-purified from nocodazole-arrested cells. The eluate was subjected to size-exclusion chromatography. Arbitrary protein abundances measured in the fractions (left) using targeted mass spectrometry²⁴ are shown for the Nup62 complex (grey; Nup62 in red) and Nup214 complex members (black). Protein abundances within fraction one (blue) are shown as a bar chart (centre; error bars indicate the standard deviation of multiple independent peptides). A silver-stained SDS-PAGE of fractions one and two is shown on the right. AU, arbitrary units. **d**, Binding of *X. laevis* Nup155 (Nup155x1), the L267D mutant, the 258–267 deletion or SUMO (negative control) to liposomes with a nuclear envelope (NE) lipid composition or 1,2-dioleoyl-*sn*-glycero-3-ethylphosphocholine (DOPC) liposomes of different sizes were analysed in flotation experiments and quantified by western blotting (columns are the average bound quantities of three independent experiments, individual data points are indicated). **e**, Scheme of the human nuclear pore scaffold architecture. NTD, N-terminal domain.

two mutant constructs affecting the aforementioned loop confirmed its importance in membrane binding (Fig. 3d). Interestingly, this orientation of Nup155 is very similar to the membrane-binding modes of Nups 160 and 133 (Extended Data Fig. 6b). We conclude that the two positions discussed above are very probably taken up by Nups 205/188 and Nup155, respectively. These interactions are probably strengthened *in situ* and challenging to detect by classical approaches when taken out of the context of the membrane. This arrangement explains the elevated copy number of Nups 205 and 93 and is consistent with a reduced copy number of Nup188 (refs 9 and 22). It is also consistent with crosslinking mass spectrometry data⁴ and FRET analysis²³ that suggested proximity to yeast Nup133 and Nup120 (Nup160 in vertebrates) *in situ*. The systematic fitting analysis suggests that the

IR is built of two slightly shifted rings on each side, similarly to the CR and NR (Fig. 3e and Extended Data Fig. 7).

Taken together, our findings point to an intriguing architectural principle in which the same basic structural building block, despite being compositionally identical, establishes locally specific conformations and interactions. Interfaces across inner and outer Y complexes appear to involve proteins that are absent in yeast, that is, Nups 43 and 358. It appears that although key features of NPC architecture, such as the Y-shaped complex itself and its head-to-tail arrangement are common to all eukaryotes, its quaternary structure might be variable across the tree of life and perhaps even across cell types.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 March; accepted 7 August 2015.

Published online 23 September 2015.

- Hoelz, A., Deblor, E. W. & Blobel, G. The structure of the nuclear pore complex. *Annu. Rev. Biochem.* **80**, 613–643 (2011).
- Vollmer, B. & Antonin, W. The diverse roles of the Nup93/Nic96 complex proteins – structural scaffolds of the nuclear pore complex with additional cellular functions. *Biol. Chem.* **395**, 515–528 (2014).
- Hurt, E. & Beck, M. Towards understanding nuclear pore complex architecture and dynamics in the age of integrative structural analysis. *Curr. Opin. Cell Biol.* **34**, 31–38 (2015).
- Bui, K. H. *et al.* Integrated structural analysis of the human nuclear pore complex scaffold. *Cell* **155**, 1233–1243 (2013).
- Kampmann, M. & Blobel, G. Three-dimensional structure and flexibility of a membrane-coating module of the nuclear pore complex. *Nature Struct. Mol. Biol.* **16**, 782–788 (2009).
- Thierbach, K. *et al.* Protein interfaces of the conserved Nup84 complex from *Chaetomium thermophilum* shown by crosslinking mass spectrometry and electron microscopy. *Structure* **21**, 1672–1682 (2013).
- Fornerod, M., van Baal, S., Valentine, V., Shapiro, D. N. & Grosveld, G. Chromosomal localization of genes encoding CAN/Nup214-interacting proteins – human CRM1 localizes to 2p16, whereas Nup88 localizes to 17p13 and is physically linked to SF2p32. *Genomics* **42**, 538–540 (1997).
- Werner, A., Flotho, A. & Melchior, F. The RanBP2/RanGAP1*SUMO1/Ubc9 complex is a multisubunit SUMO E3 ligase. *Mol. Cell* **46**, 287–298 (2012).
- Ori, A. *et al.* Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Mol. Syst. Biol.* **9**, 648 (2013).
- Stuwe, T. *et al.* Nuclear pores. Architecture of the nuclear pore complex coat. *Science* **347**, 1148–1152 (2015).
- Bilokapic, S. & Schwartz, T. U. Molecular basis for Nup37 and ELY5/ELYS recruitment to the nuclear pore complex. *Proc. Natl Acad. Sci. USA* **109**, 15241–15246 (2012).
- Seo, H. S. *et al.* Structural and functional analysis of Nup120 suggests ring formation of the Nup84 complex. *Proc. Natl Acad. Sci. USA* **106**, 14281–14286 (2009).
- Nagy, V. *et al.* Structure of a trimeric nucleoporin complex reveals alternate oligomerization states. *Proc. Natl Acad. Sci. USA* **106**, 17693–17698 (2009).
- Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).
- Boehmer, T., Jeudy, S., Berke, I. C. & Schwartz, T. U. Structural and functional studies of Nup107/Nup133 interaction and its implications for the architecture of the nuclear pore complex. *Mol. Cell* **30**, 721–731 (2008).
- Glavy, J. S. *et al.* Cell-cycle-dependent phosphorylation of the nuclear pore Nup107–160 subcomplex. *Proc. Natl Acad. Sci. USA* **104**, 3811–3816 (2007).
- Laurell, E. *et al.* Phosphorylation of Nup98 by multiple kinases is crucial for NPC disassembly during mitotic entry. *Cell* **144**, 539–550 (2011).
- Shi, Y. *et al.* Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol. Cell. Proteomics* **13**, 2927–2943 (2014).
- Culjkovic-Kraljic, B., Baguet, A., Volpon, L., Amri, A. & Borden, K. L. The oncogene *elF4E* reprograms the nuclear pore complex to promote mRNA export and oncogenic transformation. *Cell Rep.* **2**, 207–215 (2012).
- Kim, D. I. *et al.* Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc. Natl Acad. Sci. USA* **111**, E2453–E2461 (2014).
- Flemming, D. *et al.* Analysis of the yeast nucleoporin Nup188 reveals a conserved S-like structure with similarity to karyopherins. *J. Struct. Biol.* **177**, 99–105 (2012).
- Theerthagiri, G., Eisenhardt, N., Schwarz, H. & Antonin, W. The nucleoporin Nup188 controls passage of membrane proteins across the nuclear pore complex. *J. Cell Biol.* **189**, 1129–1142 (2010).
- Damelin, M. & Silver, P. A. *In situ* analysis of spatial relationships between proteins of the nuclear pore complex. *Biophys. J.* **83**, 3626–3636 (2002).
- Gaik, M. *et al.* Structural basis for assembly and function of the Nup82 complex in the nuclear pore scaffold. *J. Cell Biol.* **208**, 283–297 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to W. Baumeister and J. Plitzko for access to the electron microscopy facility of the Max Planck Institute of Biochemistry. We thank F. Schur, K. Beck, J. Briggs, C. Sachse, E. Hurt and F. Melchior for their critical advice and A. Neal for critical reading of the manuscript. We gratefully acknowledge support from EMBL's mechanical workshop, the Electron Microscopy and Proteomics Core Facilities, the Centre for Statistical Data Analysis and thank J. Krijgsveld, J. Kirkpatrick and B. Klaus. K.H.B. was supported by postdoctoral fellowships from the Swiss National Science Foundation, the European Molecular Biology Organization and Marie Curie Actions. A.O. was supported by postdoctoral fellowships from the Alexander von Humboldt Foundation and Marie Curie Actions. A.L.D. was supported by the Robert Crooks Stanley Fellowship at the Stevens Institute of Technology and National Institute on Aging (NIA) grant 1R21AG047433-01. J.K. was supported by the EMBL Interdisciplinary Postdoc Programme under Marie Curie COFUND Actions. J.S.G. was supported by an Ignition Grant Initiative from Stevens Institute of Technology and NIA grant 1R21AG047433-01. M.B. acknowledges funding by EMBL and the European Research Council (309271-NPCAtlas).

Author Contributions A.v.A., J.K. and W.A. designed and performed experiments, analysed data and wrote the manuscript. A.v.A., J.K. and P.K. performed modelling. B.V., L.S., A.O., A.L.D., M.-T.M., K.B., W.H., A.A.-P., N.B. and S.M. designed and performed experiments, and analysed data. L.P. analysed data and performed modelling. J.S.G., E.A.L. and P.B. designed experiments and oversaw the project. K.H.B. designed and performed experiments, analysed data, oversaw the project and wrote the manuscript. M.B. designed experiments, analysed data, oversaw the project and wrote the manuscript.

Author Information Associated with this manuscript are Electron Microscopy Data Bank entries EMD-3103, EMD-3104, EMD-3105, EMD-3106 and EMD-3107 and Protein Data Bank entry 5A9Q. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.H.B. (huy.bui@mcgill.ca) or M.B. (martin.beck@embl.de).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Molecular cloning and cell culture. HeLa FLP-In T-REX cell line²⁵ was stably transfected with an inducible plasmid expressing EmGFP and a microRNA against the ORF of the Nup358 mRNA using the BLOCK-iT Inducible Pol II miRNA RNAi Expression Vector Kit with EmGFP from Life Technologies (with a modified Gateway destination vector compatible with the FLP-In system⁴). Cells were treated for 4 days with $1 \mu\text{g ml}^{-1}$ of tetracycline to induce the expression of EmGFP and the miRNA. Isolation and plunge freezing of nuclear envelopes was carried out as previously described⁴. All cell lines used in this study have been regularly tested for *Mycoplasma* contamination but have not been authenticated.

Nup188, Nup205, Nup93, Nup155, Nup85 and Nup62 cDNAs were purchased from the human ORFeome collection, subcloned into a Gateway destination vector with an N-terminal His6-HA-StrepII-tag, and stably transfected into the cell line 293 FLP-In T-REX (Life Technologies). Cells were treated for 8 days with $1 \mu\text{g ml}^{-1}$ of tetracycline to induce the expression of the fusion protein, which was subsequently affinity purified.

Nuclear transport assays. HeLa cells of the Nup358-knockdown (KD) and control condition were washed twice with transport buffer (20 mM HEPES, 110 mM KOAc, 5 mM NaOAc, 2 mM MgOAc, 1 mM EGTA, pH 7.3), incubated for 15 min with Hoechst 33342 ($1 \mu\text{g ml}^{-1}$), permeabilized via incubation for 5 min on ice with digitonin ($40 \mu\text{g ml}^{-1}$), washed two times with transport buffer supplied with 1.5% polyvinylpyrrolidone (PVP, 360 kDa) to remove the soluble factors and digitonin before the addition of the transport mix. The transport mix consists of the purified transporters and recycling factors with the addition of a source of energy (1 μM Imp β , 1 μM Imp α , 4 μM Ran, 2 μM NTF2, 0.5 μM NLS-MBP-GFP, 2 mM GTP, 1 mM DTT in transport buffer).

Affinity purification and mass spectrometric analysis. Protein expression and affinity purification was performed as described before⁴. In brief, Hek-293 FLP Trex cells expressing N-terminally His6-HA-StrepII-tagged bait protein were grown in sixteen 245×245 mm plates (Nunc) per experiment. In case of Nup188 and Nup62, the cells were arrested with 200 ng ml^{-1} nocodazole (Calbiochem) 18 h before collection and freezing in liquid nitrogen. Cells expressing Nup155, Nup93 and Nup85 were collected without nocodazole arrest. Cells expressing Nup205 were applied to both approaches. Sixteen plates resulted typically in 5–7 ml of cell pellet. The cells were thawed on ice and lysed subsequently in 2.5 pellet volumes. For cells that were not nocodazole arrested, the cell suspension was sonicated 10 times for 30 s to solubilize proteins; each sonication cycle was followed by 30 s incubation. The cleared lysate was split into four and each part was passed twice over a Strep-Tactin Sepharose (IBA) column with a 400 μl bed volume by gravity flow at room temperature. After washing, the proteins were eluted with 2 mM Biotin (Sigma) in three elution steps per column ($2 \times 350 \mu\text{l}$, $1 \times 200 \mu\text{l}$). The eluate was combined and concentrated 60-fold to 60 μl final volume with (Millipore, 100 kDa molecular weight cut-off). The yield was typically between 30–160 μg of protein for cells expressing affinity-tagged Nup205, Nup188 and Nup62 with nocodazole arrest and between 30 and 100 μg for cells expressing affinity-tagged Nup205, Nup155, Nup93 and Nup85 without treatment (here, sonication was used for protein solubilization). Up to 50 μl of the concentrated eluate were subjected to size-exclusion chromatography using a Superose-6 column (GE Healthcare, 2.4 ml volume). The eluting protein was collected in 100- μl fractions. Fractions were analysed with targeted proteomics as previously described^{24,26}. In short, AQUA peptides were spiked into each fraction in equimolar amount before protein digestion. To avoid over-digestion and loss of peptides during the digestion procedure, 5 μg of heat-inactivated, cleared and denatured yeast lysate ($2 \mu\text{g l}^{-1}$ in 10 M Urea) were added to each fraction before digestion. The summed intensity of all transitions per protein resulting from mass spectrometric data acquisition were averaged and normalized to Nup93 in case of Nup62 and Nup188, and to the bait in case of other proteins to result in the final protein abundance. Error bars indicate ± 1 s.d. between the intensities of independent AQUA peptides measured for the same protein.

Mass spectrometric analysis of the Nup358 knockdown was performed according to established standards in the field of proteomics, namely three biological replicates each of the Nup358 knockdown and control condition were analysed. The shotgun proteomics data shown in Extended Data Fig. 5c were acquired as previously described⁴; *P* values were calculated using a two-sided *t*-test with empirical Bayes correction as implemented in limma. The targeted proteomic measurements shown in Fig. 2e were carried out as previously described²⁶; the *P* value given in the legend of Fig. 2e was calculated using a one-sided Welch *t*-test combined using Fisher's method. All the targeted proteomic data were analysed using SpectroDive (Biognosys AG).

Mapping of Nup155 membrane-binding motif. *Xenopus* Nup155 was cloned as a codon-optimized (for expression in *E. coli*, Genent) cDNA into a modified pET28a vector with a yeast SUMO solubility tag followed by a TEV site. The corresponding L258D mutant and the 258–267 deletion was generated by mutagenesis and deletion mutagenesis, respectively, using QuikChange site-directed mutagenesis kit (Agilent). Proteins were expressed in *E. coli* and purified using Ni-agarose. His6 and SUMO tags were cleaved using TEV protease and proteins were concentrated using VIVASPIN columns (Sartorius) and separated by gel filtration (Superdex200 PC 3.2/30, GE Healthcare) in HEPES buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 1 mM DTT). SUMO was expressed and purified from the corresponding empty vectors.

NE lipid mixture (60 mol% 1- α -phosphatidylcholine, 19.8 mol% 1- α -phosphatidylethanolamine, 10 mol% 1- α -phosphatidylinositol, 5 mol% cholesterol, 2.5 mol% sphingomyelin, 2.5 mol% 1- α -phosphatidylserine, 0.2 mol% 18:1-12:0 NBD-PE (1-oleoyl-2-[12-[(7-nitro-2,1,3-benzoxadiazol-4-yl)amino]dodecanoyl]-sn-glycero-3-phosphoethanolamine)) or DOPC mixture (99.2 mol% 1,2-dioleoyl-sn-glycero-3-ethylphosphocholine (DOPC), 0.2 mol% 18:1-12:0 NBD-PE) were dissolved in chloroform to a final concentration of 1 mg ml^{-1} . Chloroform was evaporated in a glass vial under a low stream of argon until an even lipid film formed, followed by incubation under vacuum for 1–2 h. Liposomes were formed by gentle addition of HEPES buffer to a final concentration of 5 mg ml^{-1} . After 1 h of incubation at 45°C , the flask was shaken to dissolve residual lipids. After ten cycles of freeze/thawing liposomes were either snap frozen in liquid nitrogen and stored in -80°C or directly used. Different sized liposomes were formed by passing liposomes sequentially through Nuclepore Track-Etched Membranes (Whatman) with defined pore sizes (400, 100 and 30 nm) at 45°C using the Avanti Mini-Extruder until desired size was reached. For 30 nm, liposomes were incubated in a sonication bath for 5 min before final extrusion. To ensure equal concentrations of different sized liposomes, fluorescence intensity was determined after extrusion using a Molecular Imager VersaDoc MP 4000 Imaging System and ImageJ. Concentrations were adjusted by dilution.

A protein/liposome mixture was prepared with a final protein concentration of 2.5 μM and a lipid concentration of 2.5 mg ml^{-1} and incubated at 25°C for 30 min. The protein/liposome mixture was mixed 1:1 with 72% sucrose in HEPES buffer. 75 μl were added into a TLS-55 centrifuge tube and overlaid by 1,800 μl 12% sucrose in HEPES buffer and topped with 300 μl HEPES buffer. Samples were spun for 2 h at 55,000 r.p.m., 25°C . Liposome-containing top layers were collected (450 μl). Fluorescence intensities of the start protein/liposome mixture and the top fraction were determined. Collected fractions were precipitated by the method described in ref. 27. To compare different samples, pellets were resuspended in normalized volumes of sample buffer according to the determined fluorescence signal. Binding efficiency was determined by western blot analysis, comparing band intensities of start materials with collected fractions.

Electron microscopy. 170 tomograms of HeLa cell nuclear envelopes were recorded using a Titan Krios TEM (FEI, Eindhoven) operated at 300 kV and equipped with Gatan Quantum 968 energy filter and K2 summit direct electron detector. Tilt series are collected with SerialEM²⁸ using a bidirectional tilt scheme starting from 0 to -45° and then from 0 to $+60^\circ$ tilt, with 3° increment. All tilts were recorded with the same exposure time and the average total dose ranged from $100\text{--}120 \text{ e}^- \text{ \AA}^{-2}$. A parallel beam just covered the camera to minimize beam current and thus charging on the sample. To minimize stage drift the delay time after tilting was set to the maximum 15 s. Defocus was distributed between -1.9 and $-4.0 \mu\text{m}$ in $0.1\text{--}\mu\text{m}$ steps. The nominal pixel size was 3.42 \AA . Four 4K frames were collected per tilt step in counting mode (0.8 s each $\sim 0.76 \text{ e}^- \text{ \AA}^{-2}$).

In case of the Nup358-KD condition, 140 tomograms were recorded as previously described⁴. In brief, the tilt series were recorded using a Titan Krios TEM (FEI, Eindhoven) operated at 300 kV and equipped with Gatan GIF2002 energy filter and Multiscan CCD using FEI Tomo4 software. Tilt series were collected from -44 to $+60^\circ$ with 4° increment with an average total dose of $100 \text{ e}^- \text{ \AA}^{-2}$ and defoci ranged from -3.0 to $-5.0 \mu\text{m}$.

Image processing. Four frames from each projection were subjected to cross-correlation analysis and summed to generate the motion-corrected projection²⁹. The defocus of individual projections was estimated by ctfind3 and all projections were CTF corrected by phase-flipping using IMOD as previously described³⁰. After tomogram reconstruction, sub-tomograms containing 2,171 NPCs, corresponding to 17,368 asymmetric units were extracted and subjected to sub-tomogram averaging. Sub-tomogram averaging was carried out on the asymmetric unit level and independently for the CR, IR and NR regions as previously described⁴ but with the following modifications. A gold standard FSC procedure was employed on two independently aligned sets of sub-tomograms. The sub-tomograms with full dose including all projections were initially aligned. Subsequently, sub-tomograms with a dose of $80 \text{ e}^- \text{ \AA}^{-2}$ including projections from -45 to $+40^\circ$ were used for refinement. The two final structures were averaged and the average of CR, IR

and NR were sharpened with empirically determined negative B-factor of $6,000 \text{ \AA}^2$, while filtering the data to a resolution of 23.4, 22.9 and 24.1 \AA respectively. These B-factors are probably due to additional dampening of high-frequency information by beam-induced motion at higher tilts, and the inaccuracy of tomogram alignment and reconstruction. To prevent over-sharpening, a range of different B-factors were applied in order to compare the resulting tomographic maps to the respective X-ray structures filtered to 23 \AA resolution (Extended Data Fig. 2). The mask during the averaging procedure was chosen such that the averaged segment was larger than the asymmetric unit. For visualization purposes, overlapping segments were fitted into each other. For the systematic fitting approach, the segments were joined.

In case of the Nup358-KD condition, sub-tomograms containing 920 NPCs, corresponding to 7,360 asymmetric units were used for sub-tomogram averaging. The final structure has a resolution of 37.5 \AA at FSC 0.5 criteria. The cryo-ET data of the Nup358-KD condition subjected to iterative multi-reference classification³¹. Aligned asymmetric units are classified by constrained cross-correlation coefficient using structures from Nup358-KD condition and untreated HeLa cells⁴ until convergence. As a result, the majority of asymmetric subunits (7,005) were classified into class 1 (outer Y complex absent in CR), while ~5% subunits (355) were classified into class 2 (outer Y complex present in CR). The average of class 2 clearly shows intact wild-type-like structure (Extended Data Fig. 5e, left). For the generation of Extended Data Fig. 5f, class 1 particles were queried for adjacent neighbours within single NPCs using a dedicated MATLAB script. The expected frequency for adjacency of outer Y complexes in case of (hypothetical) stochastic binding was calculated using R.

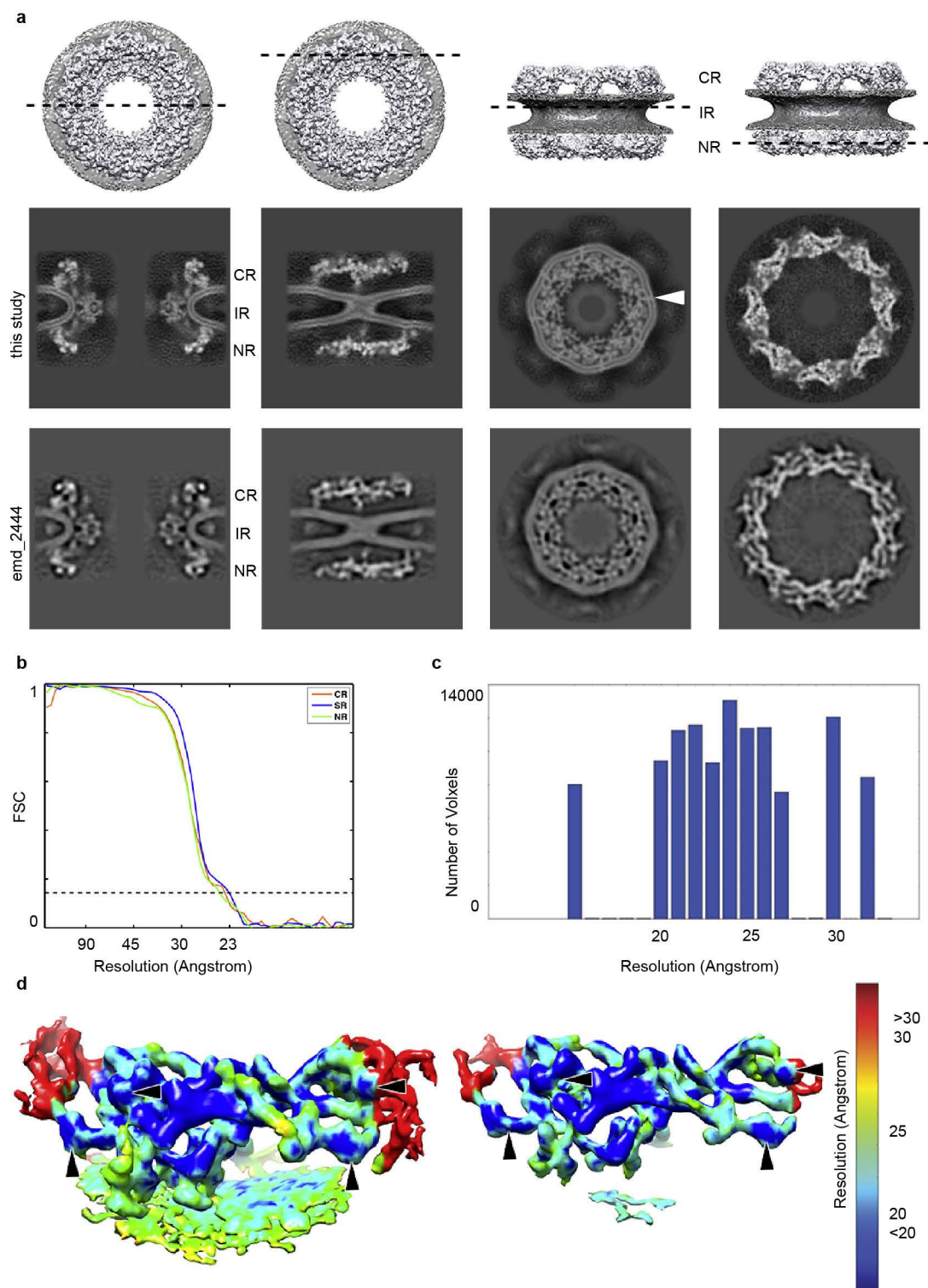
Homology modelling. MODexplorer³² was used to detect modelling templates and generate initial sequence alignments to the templates. Swiss PDB Viewer³³ was used to optimize the alignments guided by alternative alignments and secondary structure predictions from GeneSilico MetaServer³⁴. Final models were built with Modeller³⁵ based on the optimized alignments. The modelling templates or selected human crystal structures are listed in Extended Data Table 1. The models were built in oligomeric state whenever a corresponding oligomeric structure was available (for example, for the Sec13–Nup96 NTD–Nup107 CTD and Nup107 NTD–Nup133 CTD complexes).

Fitting of high-resolution structures. The models were fitted to the EM map of the joined NR, IR and CR segments using UCSF Chimera³⁶ Fit In Map tool. Each model was separately fitted using systematic global search with an arbitrarily large number of 1,000,000 random initial placements and a normalized cross-correlation score as a fitting metric. The fits were clustered with the Fit In Map tool, taking the eight-fold symmetry into account, leading to on average 20,000–60,000 representative unique fits. The fits were scored with the normalized cross-correlation and the statistical significance of the scores was assessed as previously described⁴. In brief, correlation scores were transformed to z-scores (Fisher's z-transform) and centred; subsequently, an empirical null distribution was fitted from which two-sided P values were computed. Then, the final fits were selected based on the fitting scores and/or agreement with the previously identified locations, cross-links and known interactions (see Supplementary Table 1 for detail). Nup85-CTD and Nup133 β -propeller domain were placed manually based on the locations expected from domain connectivity and vicinity to the membrane (Nup133). The fits of the Nup107–Nup133 and Nup37–Nup160 subcomplexes were apparent to be non-optimal from the rigid body fitting and thus were further optimized flexibly: (i) the Nup107–Nup133 structure was divided into two substructures (Nup107-NTD and Nup133–Nup133 dimer) based on the previously identified

hinge region³⁷ and re-fitted locally; (ii) the Nup37–Nup160 structure was subjected to the Normal Mode Analysis using ElNemo³⁸ and a model that optimally fitted Nup160-CTD to the EM density without clashing with neighbouring subunits served as a basis for separately placing the Nup160-CTD in the map. Some of the fits such as Nup155, Nup37–Nup160 NTD were further optimized locally. Interfaces of Nup43–Nup85 and Sec13–Nup107 were optimized with local re-docking or refinement, respectively, using Haddock³⁹. Although the orientation of Nup43 was not clear from the systematic fitting, the re-docking led to the best scoring model satisfying cross-links to Seh1 and Nup85-CTD.

Mapping of phosphorylation sites. The phosphorylation sites from human and mouse have been collected from the PTMcode database⁴⁰.

25. Zemp, I. *et al.* Distinct cytoplasmic maturation steps of 40S ribosomal subunit precursors require hRio2. *J. Cell Biol.* **185**, 1167–1180 (2009).
26. Ori, A., Andres-Pons, A. & Beck, M. The use of targeted proteomics to determine the stoichiometry of large macromolecular assemblies. *Methods Cell Biol.* **122**, 117–146 (2014).
27. Wessel, D. & Flugge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138**, 141–143 (1984).
28. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
29. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).
30. Schur, F. K., Hagen, W. J., de Marco, A. & Briggs, J. A. Determination of protein structure at 8.5 \AA resolution using cryo-electron tomography and sub-tomogram averaging. *J. Struct. Biol.* **184**, 394–400 (2013).
31. Movassagh, T., Bui, K. H., Sakakibara, H., Oiwa, K. & Ishikawa, T. Nucleotide-induced global conformational changes of flagellar dynein arms revealed by *in situ* analysis. *Nature Struct. Mol. Biol.* **17**, 761–767 (2010).
32. Kosinski, J., Barbato, A. & Tramontano, A. MODexplorer: an integrated tool for exploring protein sequence, structure and function relationships. *Bioinformatics* **29**, 953–954 (2013).
33. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
34. Kuroski, M. A. & Bujnicki, J. M. GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.* **31**, 3305–3307 (2003).
35. Eswar, N. *et al.* Comparative protein structure modeling using MODELLER. *Current Protoc. Protein Sci.* Ch. 2 (2007).
36. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
37. Whittle, J. R. & Schwartz, T. U. Architectural nucleoporins Nup157/170 and Nup133 are structurally related and descend from a second ancestral element. *J. Biol. Chem.* **284**, 28442–28452 (2009).
38. Suhre, K. & Sanejouand, Y. H. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.* **32**, W610–W614 (2004).
39. de Vries, S. J. *et al.* HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* **69**, 726–733 (2007).
40. Minguez, P. *et al.* PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic Acids Res.* **43**, D494–D502 (2015).
41. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11**, 63–65 (2014).
42. Walther, T. C. *et al.* The cytoplasmic filaments of the nuclear pore complex are dispensable for selective nuclear protein import. *J. Cell Biol.* **158**, 63–77 (2002).
43. Drin, G. *et al.* A general amphipathic α -helical motif for sensing membrane curvature. *Nature Struct. Mol. Biol.* **14**, 138–146 (2007).

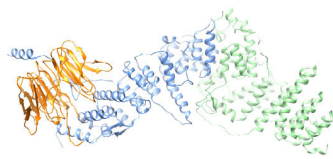


Extended Data Figure 1 | Tomographic map of the human NPC.

a, Orthoslices through the nucleocytoplasmic axis, CR, IR and NR of a tomographic structure of the human NPC obtained using a direct electron detector (this study) compared to a structure obtained using a conventional detector using a similar experimental workflow (Electron Microscopy Database accession number EMD-2444). In both cases, the CR, IR and NR were aligned independently. The arrowhead indicates a transmembrane domain that is resolved in the IR. **b**, Fourier shell correlation curves of the CR, IR and NR regions. **c**, Histogram corresponding to the colour-coded local resolution map

shown in **d** that was calculated using ResMap⁴¹. **d**, A single segment of the NR ring is shown (in all other figures the segments are shown jointly with their anterior and/or posterior asymmetric unit) at two different isosurface thresholds. The redundant density of the outer Nup43 β -propeller (horizontal arrowhead) and the Nup133 middle domain (vertical arrowhead) of two neighbouring asymmetric units are indicated for orientation. The reduced resolution at the edges is due to the border of the mask used during alignment and averaging that covered about 1.5 times the asymmetric unit.

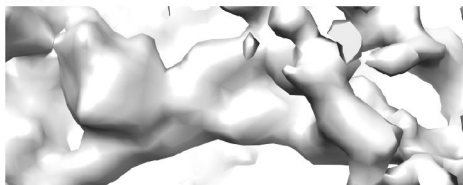
a Nup107-NTD/Nup96/Sec13



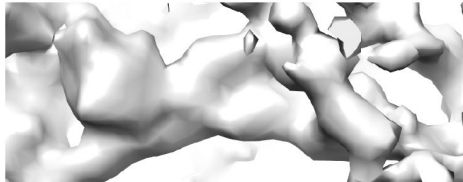
filtered 23 A



9000 A²



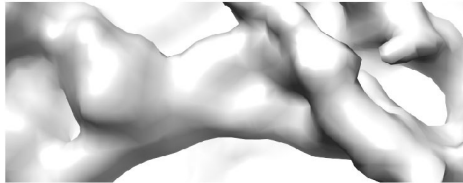
8000 A²



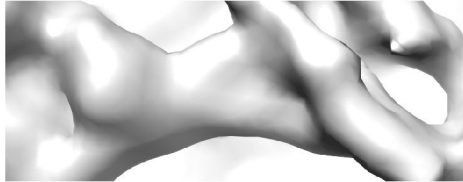
7000 A²



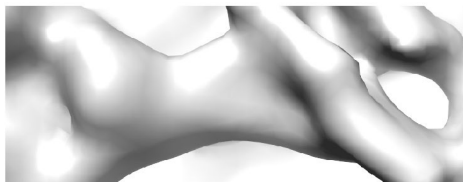
6000 A²



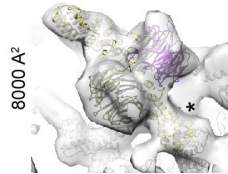
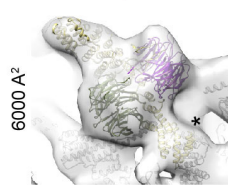
4000 A²



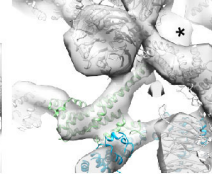
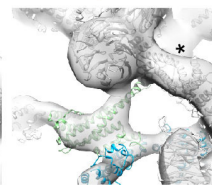
2000 A²



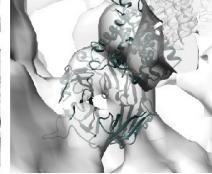
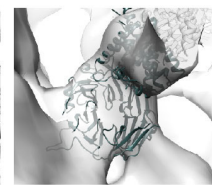
b Nup85-NTD/
Seh1/Nup43



Nup107-NTD/
Nup133

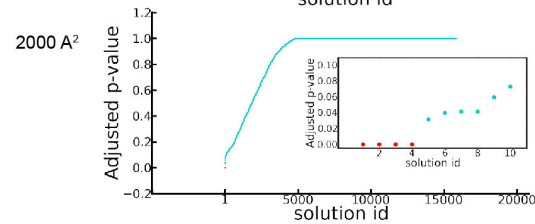
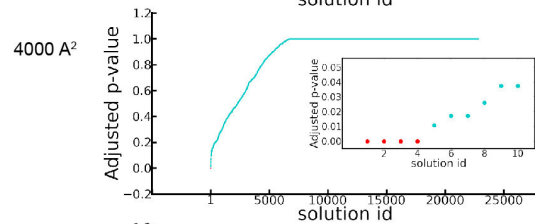
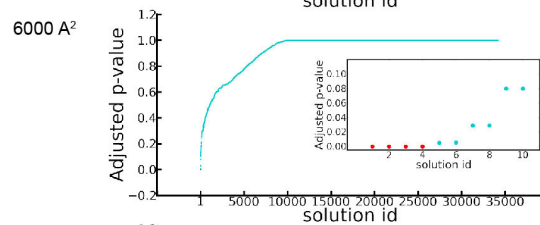
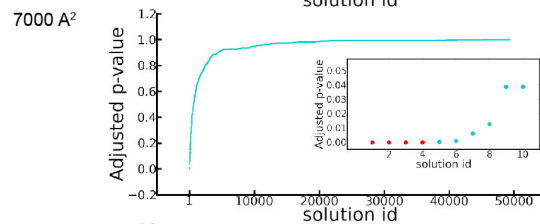
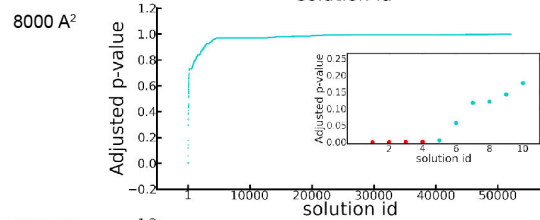
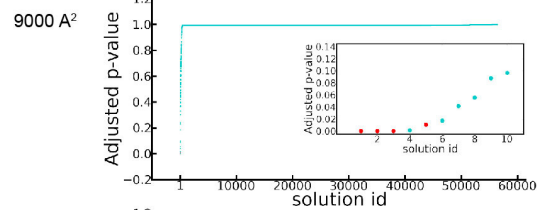


Nup155-NTD



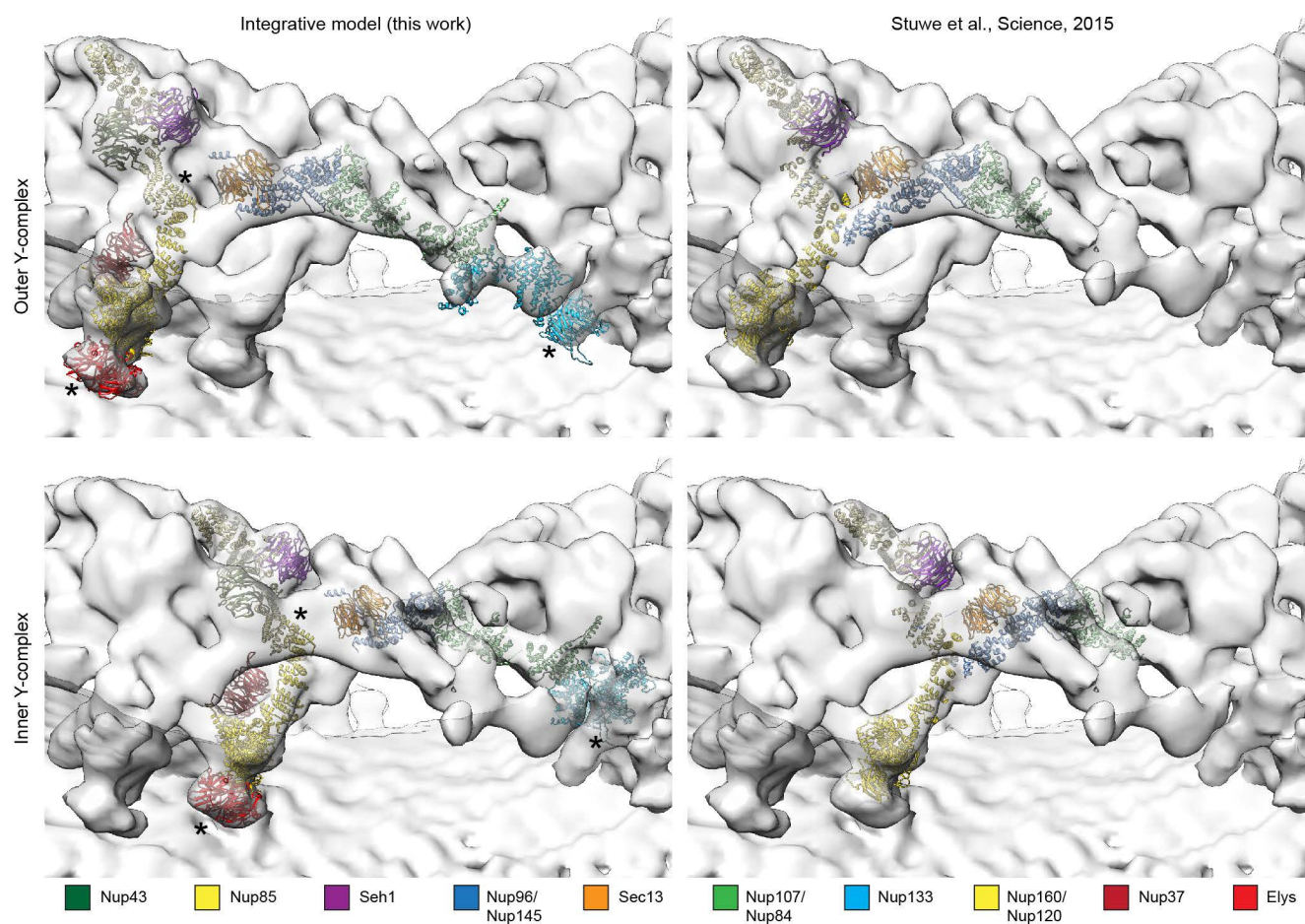
c

Nup107-NTD/Nup96/Sec13



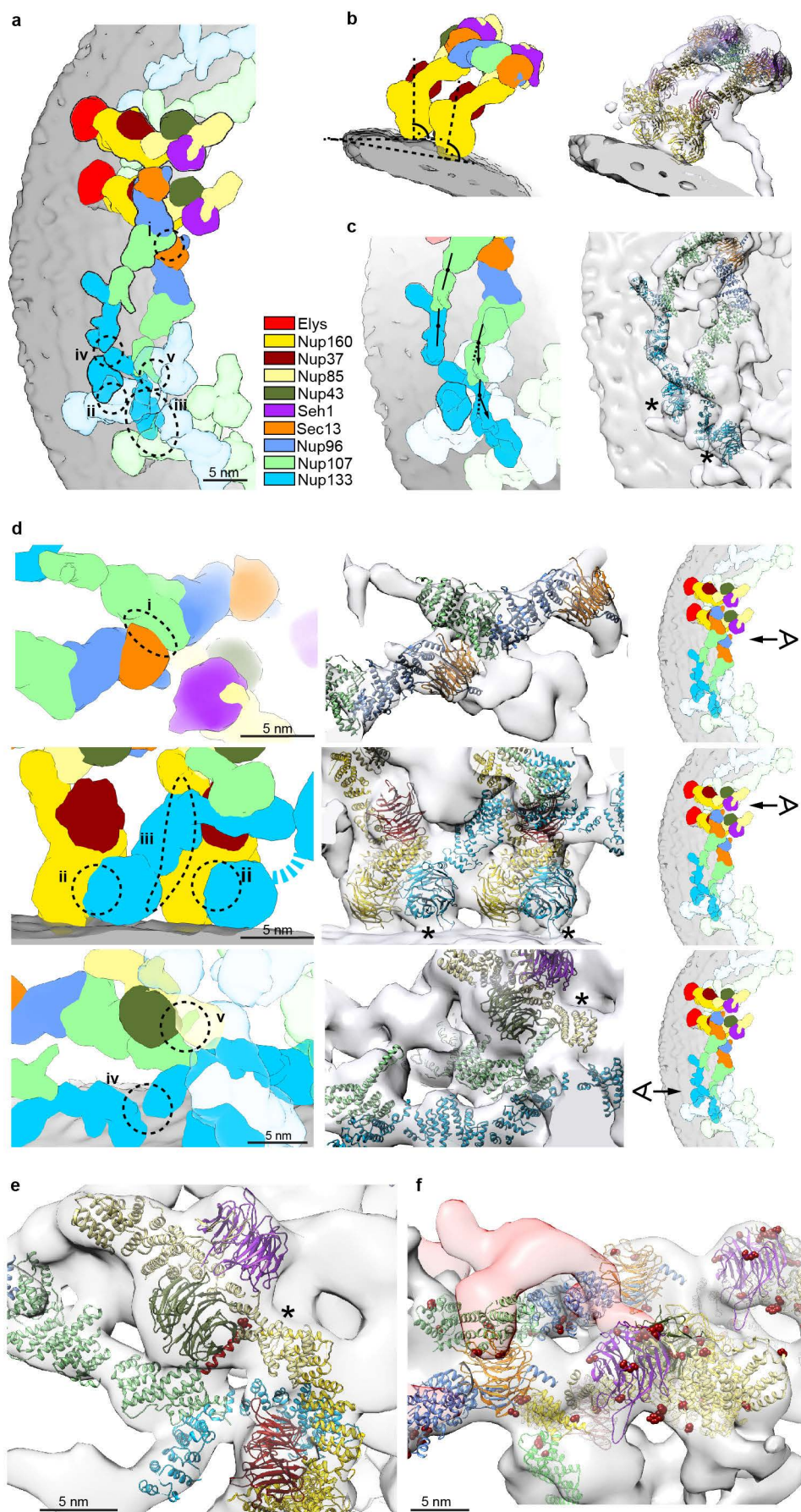
Extended Data Figure 2 | B-factor correction. X-ray structures filtered to the overall resolution of the tomographic map were compared to tomographic maps corrected with different B-factors to choose an appropriate B-factor. **a**, X-ray structure of Nup107–Nup96–Sec13 (top), filtered to 23 Å (below) as compared to respective regions of the outer vertex corrected with B-factors ranging from 2,000–9,000 Å². **b**, B-factors of 6,000–8,000 Å² most realistically resemble features of the X-ray structures. Three regions of the tomographic map are superimposed with the respective X-ray structures at B-factors of 6,000 Å² in comparison to 8,000 Å². In these well-resolved regions, additional features such as more detailed shapes of β-propellers or the Nup107 finger

domain (see also Fig. 2a, b and Extended Data Fig. 4e) are apparent at 8,000 Å². Owing to local deviations in resolution (Extended Data Fig. 1c, d) a more conservative B-factor of 6,000 Å² was chosen to correct the averages. Asterisks indicate structures that can be unambiguously positioned but have some uncertainty in their orientation, that is the Nup85 carboxy-terminal domain (CTD). **c**, Systematic fitting of the X-ray structure of Nup107–Nup96–Sec13 into the tomographic map as shown in Extended Data Fig. 7a but at different B-factors. Adjusted *P* values are shown ranked; the four true positive hits are shown in red in the inset. The latter are consistently identified as top hits, except when a B-factor of 9,000 Å² is used.



Extended Data Figure 3 | Comparison of hybrid model of the Y complex to the X-ray structure of the vertex region¹⁰. The hybrid model of the Y complex (NR) shown on the left side was generated independently from the coordinates X-ray structure of the vertex shown on the right side. The structures of Y complex members were fitted into the tomographic map based on spatial restraints and complementary information (see Supplementary

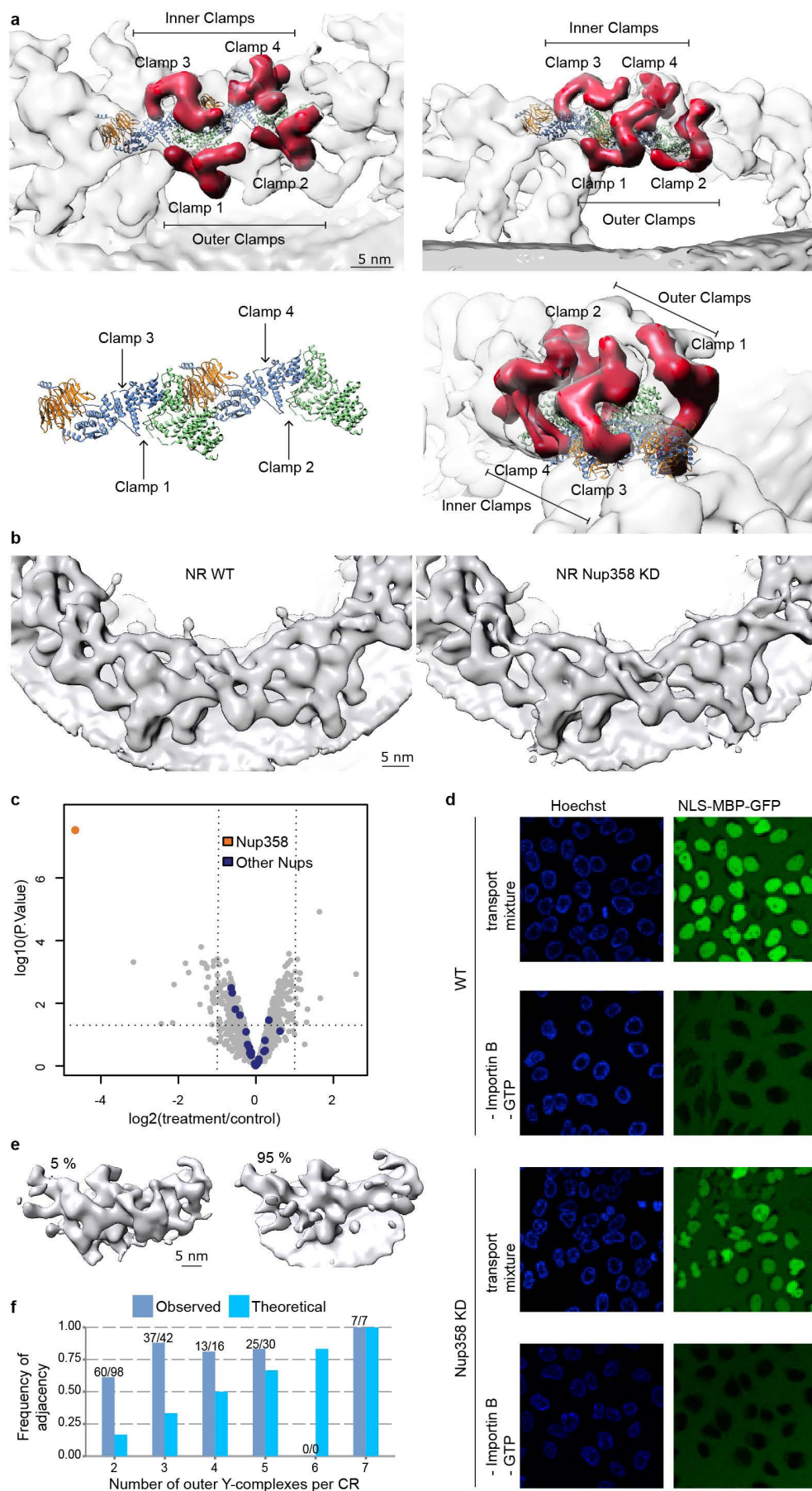
Table 1 for details). The outer and inner Y complexes/vertices are shown separately on the top and the bottom. The molecular weight of one human Y complex is approximately 1 MDa, the majority of which is structured. Asterisks indicate structures that can be unambiguously positioned but have some uncertainty in their orientation. Those are the β -propellers of Elys and Nup133 as well as Nup85-CTD.



Extended Data Figure 4 | The inner and outer Y complexes have distinct conformations and engage in locally specific sets of interactions.

a, Arrangement of the inner and outer Y complex as seen from above. X-ray structures were filtered to 2.3 nm resolution and coloured by protein. Positions of the five interfaces between Y complexes are indicated. **b**, The inner and outer copies of Nup160 assume the same normal vector with respect to the membrane and are slightly tilted to each other because of the different diameters from the central axis. **c**, The hinges between the Nup107 C- and N-terminal domains as well as within the Nup133 C-terminal domain have a different conformation in the inner and outer Y complex. In the case of the inner stem, the middle domain of Nup133 appears slightly bent inwards compared to the conformation revealed in the X-ray structure, which can be accounted for by introducing a hinge, as previously predicted based on the Nup133 structure³⁷. **d**, Magnified views showing details of five interfaces (I–V). The panels on the right indicate the viewing point. (I) Sec13 of the outer vertex interfaces with the Nup107 N-terminal domain of the inner vertex^{4,10}; (II) the N-terminal β -propeller of Nup133 interfaces with Nup160 of the posterior asymmetric unit in the case of both Y complexes, forming the head-to-tail contact that facilitates ring formation^{12,14}; (III) therefore, only the β -propeller of

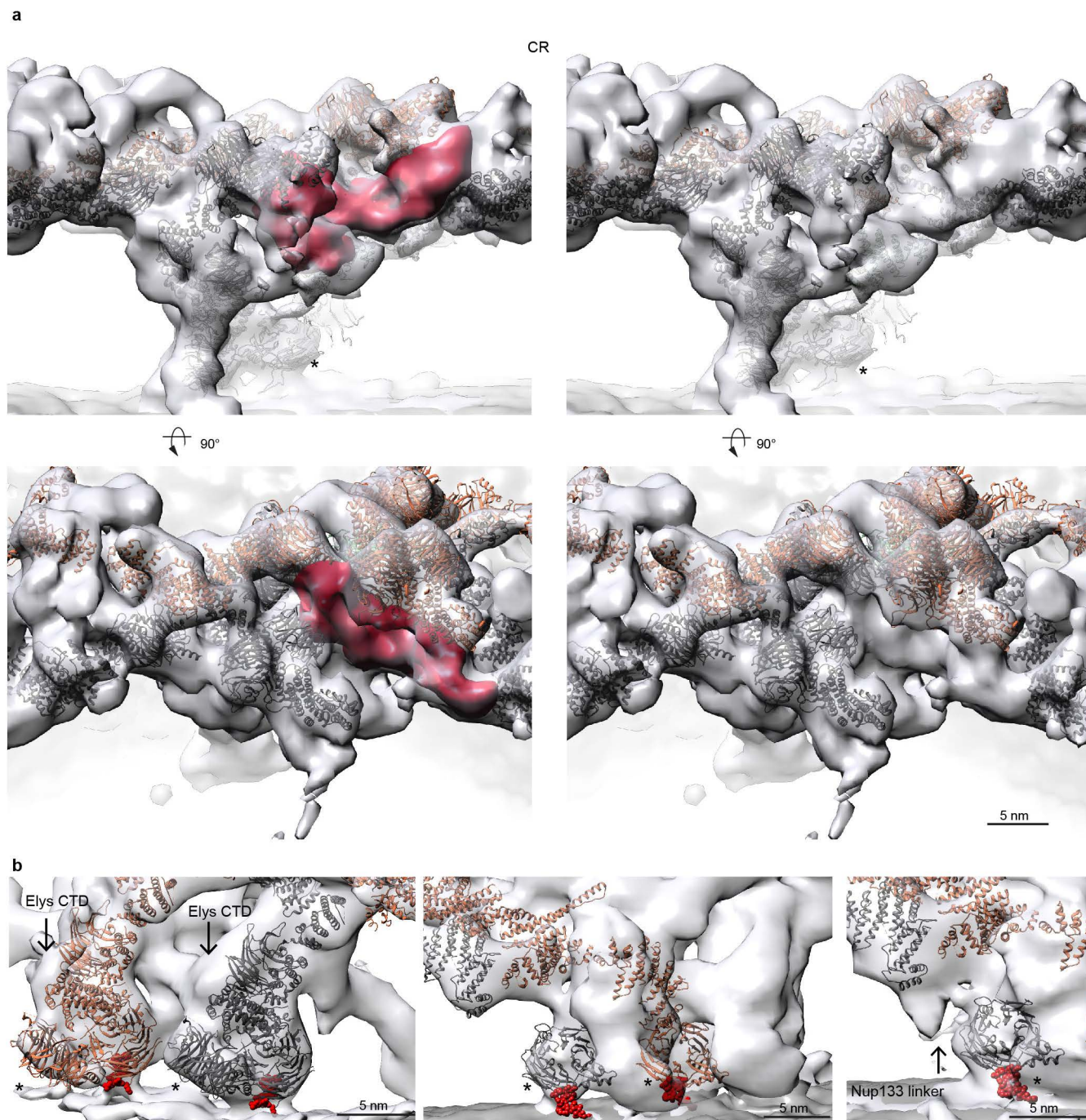
the inner copy of Nup133 is sandwiched between both Nup160 proteins. In this case, the N-terminal α -helical domain extends into a larger interface with the outer Nup160 of the posterior asymmetric unit; (IV) the C terminus of the inner Nup133 branches out of the inner stem to form a contact with its counterpart on the outer stem. This relatively small interface is reminiscent of a crystal contact observed in the Nup133–Nup107 structure (Protein Data Bank accession number 3I4R)³⁷; and (V) Nups 85 and 43 of the outer vertex form an interface with the C-terminal domain of Nup107 of the inner stem. **e**, The Nup107 finger domain (red) is exclusive to higher eukaryotes. While its inner copy (shown) interfaces with Nup43 and Nup85 of the outer vertex, its outer copy interfaces with the density connecting both Y complexes (Fig. 2a, b). A phosphorylation site in the finger domain is depicted as dark-red spheres. Asterisks indicate structures that can be unambiguously positioned but have some uncertainty in their orientation, that is Nup85-CTD. **f**, Vertex region of the CR as seen from the central channel. Phosphorylation sites are represented as in **e**. The density connecting both Y complexes is segmented as in Fig. 2b and is in close contact with several phosphorylation sites in Sec13, Nup96 and Nup107. Phosphorylation sites in Seh1 and Nup85 are in direct proximity to the Nup214 complex region.



Extended Data Figure 5 | Connection of the inner and outer Y complexes.

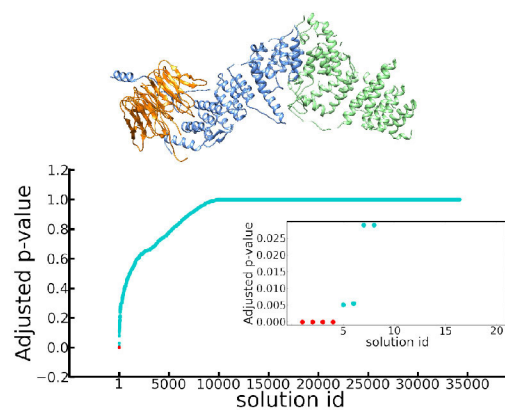
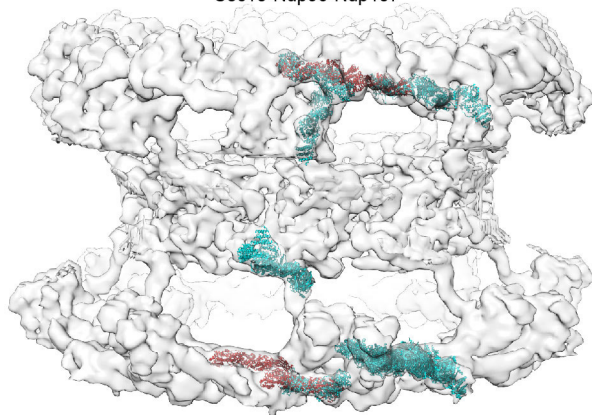
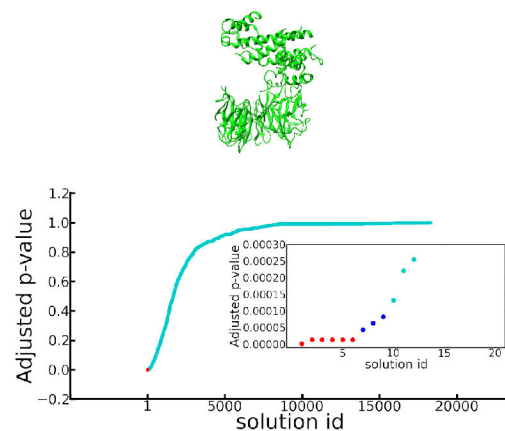
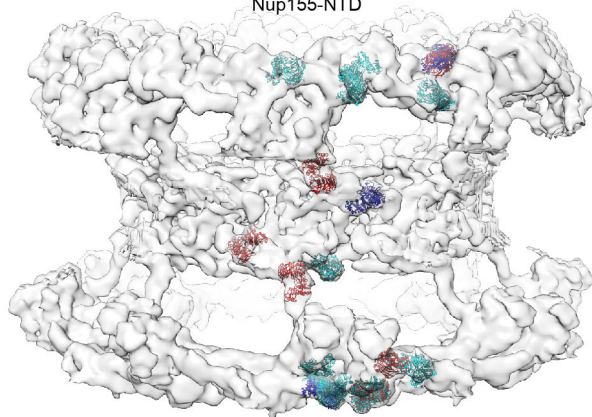
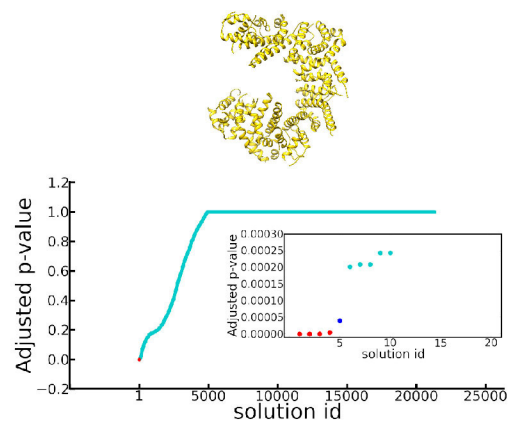
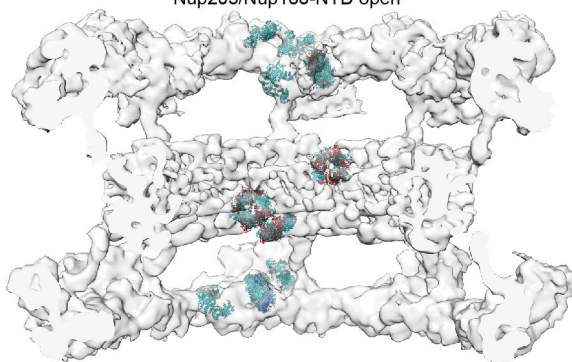
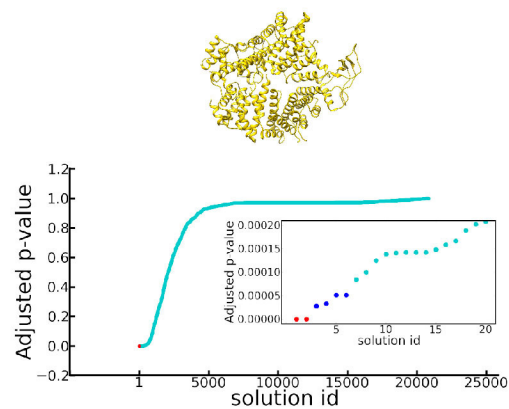
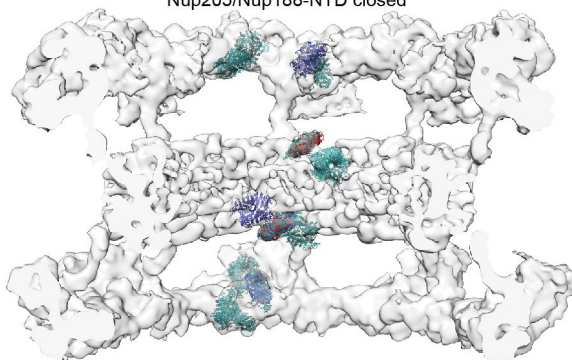
a, Four clamp-shaped densities (segmented red) emanate from the Nup96/107 region of both Y complexes in the CR. Only the inner clamps connect both Y complexes, whereas the outer ones protrude into a more complex substructure at the outer periphery of the CR (Fig. 2b). **b**, Same as in Fig. 2c, d but for the NR. **c**, Volcano plot visualizing shotgun proteomics data obtained of HeLa cells in the Nup358 knockdown (treatment) as compared to the control condition. **d**, Nuclear transport assays of NLS-MBP-GFP (that is, a nuclear localization sequence bound to a GFP-tagged maltose-binding protein) in non-treated cells and the Nup358 knockdown condition. Cargos with a classical nuclear localization signal are imported in the absence of Nup358 as previously observed⁴², although a lower efficiency cannot be excluded. **e**, Classification of sub-tomograms of the knockdown condition reveals that approximately 5% of all asymmetric units contain an outer Y complex in the CR, which is in excellent agreement with the knockdown efficiency of ~95%. Classification was done on the level of asymmetric units. Transferred to the level of NPCs, it suggests that out of 920 NPCs observed under gene silencing

conditions, 663 had no outer Y complex in the CR, 183 had one, 49 had two, 14 had three, 4 had four, 6 had five, 0 had six, 1 had seven and 0 had eight. **f**, The observed adjacency of outer Y complexes in the CR under knockdown conditions was much higher than expected. The 5% of asymmetric units that contained outer Y complexes in the CR were analysed on the NPC level to determine whether their neighbouring asymmetric units also contained outer Y complexes in the CR. The observed frequency of adjacency is shown in dark blue. The respective total number of observed outer Y complexes in the CR and the number of the ones that had adjacent partners is indicated as (n/m). The observed frequency is considerably elevated over the theoretical frequency (shown in bright blue) that would be expected if Y complexes would bind to random subunits. This observation implies cooperativity for Y complex assembly/maintenance within the CR that might arise through the head-to-tail contact of adjacent Y complexes. NPCs with zero, one and eight outer Y complexes per CR are not shown because they cannot contain any adjacent Y complexes or were not observed.



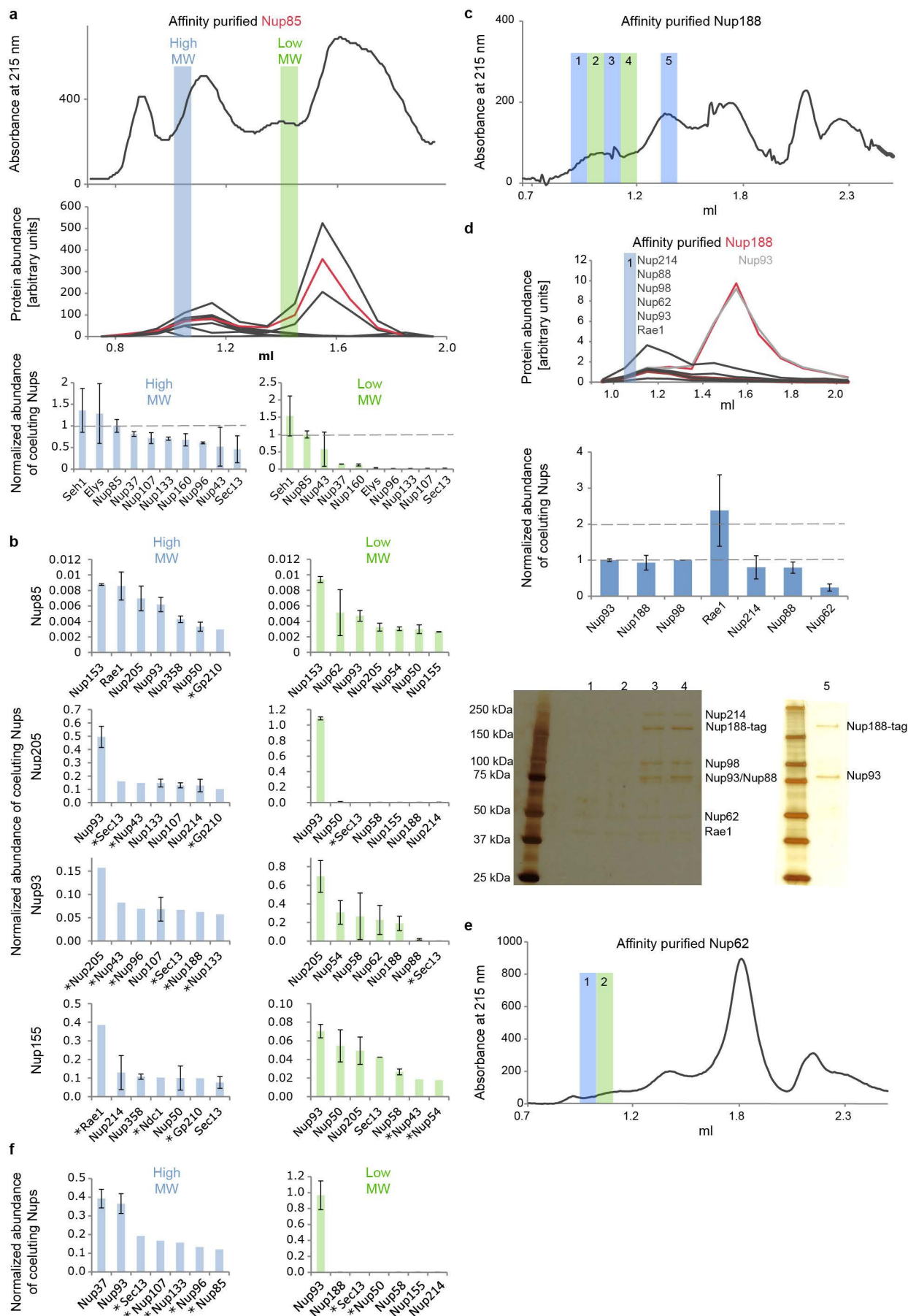
Extended Data Figure 6 | Structural signatures of inner-ring scaffold Nups and membrane-binding motifs of Nup160 and Nup133. a, Same as Fig. 3a but for the CR. **b,** Fits of Nup160 (left) of the outer (orange) and inner (grey) Y complexes into the NR are shown. Additional density accounting for the C-terminal domain of Elys is indicated. Fits of Nup133 are shown at normal (centre) and high (right) isosurface thresholds. At the high isosurface

threshold, density linking both Nup133 domains is also apparent in the outer stem. The membrane-binding motifs⁴³ are coloured red. Asterisks indicate structures that can be unambiguously positioned but have some uncertainty in their orientation. Those are the β -propellers of Elys and Nup133 as well as Nup85-CTD.

a Sec13-Nup96-Nup107**b** Nup155-NTD**c** Nup205/Nup188-NTD open**d** Nup205/Nup188-NTD closed

Extended Data Figure 7 | Systematic fitting of selected NPC components to the electron microscopy map. Each panel shows the 20 best-scoring fits (left), a plot of P values for all the solutions (right) and the top solutions in the inset. The models used for fitting are shown as ribbon representation. The fits and data points are coloured according to the groups with similar P value ranges. The group of fits with the best P values is coloured red (high-confidence fits), the second-best group (medium-confidence fits) is coloured blue, and

all remaining are coloured cyan. The membrane density has been removed for clarity. **a**, The Sec13–Nup96–Nup107 subcomplex, for which the ground-truth is known because it is part of the vertex. **b–d**, same as **a** but for the N-terminal domains of Nup155, the open conformation of Nup205/188-NTD (template Nup205) and the closed conformation of Nup205/188-NTD (template Nup188), respectively. Solution id, solution identity.



Extended Data Figure 8 | Co-elution analysis to detect weak nucleoporin interactions. To detect weak interactions of scaffold nucleoporins we combined rapid affinity isolation with gel filtration and quantitative targeted proteomics to measure absolute protein abundances. HEK293 cells expressing various affinity-tagged Nups (in contrast to Fig. 3c without nocodazole arrest) were lysed using mild conditions and sonication for protein solubilization. Affinity isolates were subjected to gel-filtration and all fractions were analysed using targeted mass spectrometry, as previously described^{24,26}, to measure protein abundances in the high- and low-molecular-weight fractions. The high-molecular-weight fractions are indicative of potential outgoing interactions of large molecular species. The low-molecular-weight fraction will highlight smaller fragments that occur after sonication. **a**, In the case of affinity-tagged Nup85 the top panel shows the 215 nm absorption curve of the gel-filtration experiment. The middle panel shows the arbitrary protein abundance units of Y-complex members in all fractions (red for Nup85, black for all other Y-complex members). Protein abundances (normalized to the affinity-tagged protein) in the high-molecular-weight fractions (blue bar) and low-molecular-weight fraction (green bar) are shown as bar charts in the bottom panel. The low-molecular-weight peak corresponds to the small arm proteins (Nup85, Seh1 and Nup43), the high-molecular-weight peak to the intact Y complex. **b**, The same approach was applied to Nup205, Nup93 and Nup155. The seven

most abundant co-eluting Nups are shown for the high-molecular-weight fractions (blue bar plots) and low-molecular-weight fractions (green bar plots). In case of Nup85 (top) the seven most abundant proteins apart from the Y-complex members are shown. Weak interactions of Y-complex members with Nup205 and Nup93 are apparent. In the case of Nup155, weak interactions are detected with CR and NR members, as well as Sec13 that localizes to the proximity of the C-terminal domain of Nup155 when fitted into the density connecting the IR with CR/NR. Tpr was excluded from this analysis since it was present in all fractions. Protein abundances based on single reference peptides are marked with an asterisk. **c**, Same as **a** (top panel) but for Nup188 affinity purified from nocodazole-arrested cells. **d**, Same as Fig. 3c but for Nup188 affinity purified from nocodazole-arrested cells. Co-eluting species in the high-molecular-weight fraction are similar to the ones observed for affinity-purified Nup62 (Fig. 3c). Isostoichiometric amounts of Nup188, Nup98, Nup93, Nup62 and an enrichment of Nup214, Nup88 and Rae1 were detected. The Nup188–Nup93 heterodimer thus binds to Nups that are well-established components of the CR, which is consistent with the systematic fitting approach. **e**, Same as **c** but corresponding to Fig. 3c. **f**, Same as **b** (second panel for Nup205) but for nocodazole-arrested cells. In the case of Nup205, the co-purifying species are similar in nocodazole-arrested as compared to untreated cells.

Extended Data Table 1 | Template structures used for homology modelling or selected human crystal structures

| Nucleoporin | AA covered by the model | Modeling template or human structure | Organism* | PDB Identifier | Model confidence† |
|-------------|-------------------------|--------------------------------------|-----------|----------------------------|-------------------|
| Nup107 | NTD (150-602) | Nup84/Nup145C/Sec13 | S.c. | 3IKO chain C | medium |
| | CTD (667-924) | Crystal structure Nup107/Nup133 | H.s. | 3I4R chain A | high |
| Nup133 | NTD (75-477) | Nup133 | H.s. | 1XKS | high |
| | CTD (518-543) | Crystal structure Nup107/Nup133 | H.s. | 3I4R chain B | high |
| Nup96 | 277-751 | Nup84/Nup145C/Sec13 | S.c. | 3IKO chain B | medium |
| Sec13 | 14-304 | Sec13/Nup145C | H.s. | 3BG1 chain A | high |
| Nup85 | NTD (9-475) | Nup85 | S.c. | 3F3F chain C | medium |
| | CTD (476-641) | Nic96 | S.c. | 2QX5 chain A | low |
| Seh1 | 1-322 | Seh1 | S.c. | 3F3F chain A | high |
| Nup37 | 9-326 | Nup37 | S.p. | 4GQ2 chain P | medium |
| Nup160 | 41-1201 | Nup120 | S.p. | 4GQ2 chain M, 4FHN chain B | medium |
| Nup43 | 4-380 | Crystal structure | H.s. | 4I79 chain A | high |
| Elys | NTD (3-493) | Elys | M.m. | 4I0O chain A | high |
| Nup93 | CTD (173-815) | Nic96 | S.c. | 2RFO chain A | medium |
| Nup188 | NTD (1-939) | Nup188 | M.t. | 4KF7 chain A | medium |
| Nup205 | NTD (9-1008) | Nup192 | S.c. | 4IFQ chain A | medium |
| | | Nup188 | M.t. | 4KF7 chain A | |
| Nup155 | NTD (20-863) | Nup157 | S.c. | 4MHC chain A | high |
| | CTD (871-1375) | Nup170 | S.c. | 3I5P chain A | high |

*Organism: H.s., *Homo sapiens*; M.m., *Mus musculus*; S.c., *Saccharomyces cerevisiae*; S.p., *Saccharomyces pombe*; M.t., *Myceliophthora thermophila*. AA, amino acid.

†Model confidence was assessed as follows: high, human crystal structures or models based on very similar templates; medium, good quality models with confident overall sequence alignment but containing less confident regions; low, models expected to have correct fold but built based on low-confidence sequence alignment.

CORRIGENDUM

doi:10.1038/nature14624

Corrigendum: Fatty acid carbon is essential for dNTP synthesis in endothelial cells

Sandra Schoors, Ulrike Bruning, Rindert Missiaen, Karla C. S. Queiroz, Gitte Borgers, Ilaria Elia, Annalisa Zecchin, Anna Rita Cantelmo, Stefan Christen, Jermaine Goveia, Ward Heggermont, Lucica Goddë, Stefan Vinckier, Paul P. Van Veldhoven, Guy Eelen, Luc Schoonjans, Holger Gerhardt, Mieke Dewerchin, Myriam Baes, Katrien De Bock, Bart Ghesquière, Sophia Y. Lunt, Sarah-Maria Fendt & Peter Carmeliet

Nature **520**, 192–197 (2015); doi:10.1038/nature14362

We thank our colleagues from the metabolism community (Emile Van Schaftingen and Guido Bommer, University of Louvain, Belgium, and Frans Schuit, University of Leuven, Belgium), who alerted us to a possible confusion arising from our Article. In particular, the statement in the abstract that “Isotope labelling studies in control endothelial cells showed that fatty acid carbons substantially replenished the Krebs cycle” and similar phrases later in the text could be misunderstood as implying anaplerosis. By no means did we intend to suggest that the Krebs cycle is replenished by the net contribution of fatty acid carbons in the traditional sense of anaplerosis. Indeed, even though labelled acetyl-CoA from fatty acids enters the Krebs cycle and labels oxaloacetate, we did not want to imply net formation of oxaloacetate from acetyl-CoA, derived from fatty acids. Rather, our data suggest that under conditions of adequate availability of anaplerotic substrates (glucose, glutamine), fatty acid oxidation is important to produce sufficient amounts of dNTPs. Fatty acids provide acetyl-CoA, which helps to sustain the TCA cycle and dNTP synthesis for proliferation in conjunction with an anaplerotic substrate. Why the anaplerotic nutrients glucose and glutamine do not sustain sufficient aspartate and nucleotide synthesis in the CPT1A-silenced endothelial cells is an intriguing but outstanding question.

CORRIGENDUM

doi:10.1038/nature14959

Corrigendum: Progesterone receptor modulates ER α action in breast cancer

Hisham Mohammed, I. Alasdair Russell, Rory Stark, Oscar M. Rueda, Theresa E. Hickey, Gerard A. Tarulli, Aurelien A. Serandour, Stephen N. Birrell, Alejandra Bruna, Amel Saadi, Suraj Menon, James Hadfield, Michelle Pugh, Ganesh V. Raj, Gordon D. Brown, Clive D'Santos, Jessica L. L. Robinson, Grace Silva, Rosalind Launchbury, Charles M. Perou, John Stingl, Carlos Caldas, Wayne D. Tilley & Jason S. Carroll

Nature **523**, 313–317 (2015); doi:10.1038/nature14583

In this Article, author Aurelien A. Serandour should have been listed with one middle initial only. This has been corrected in the online versions.

CORRIGENDUM

doi:10.1038/nature14960

Corrigendum: Bipolar seesaw control on last interglacial sea level

G. Marino, E. J. Rohling, L. Rodríguez-Sanz, K. M. Grant, D. Heslop, A. P. Roberts, J. D. Stanford & J. Yu

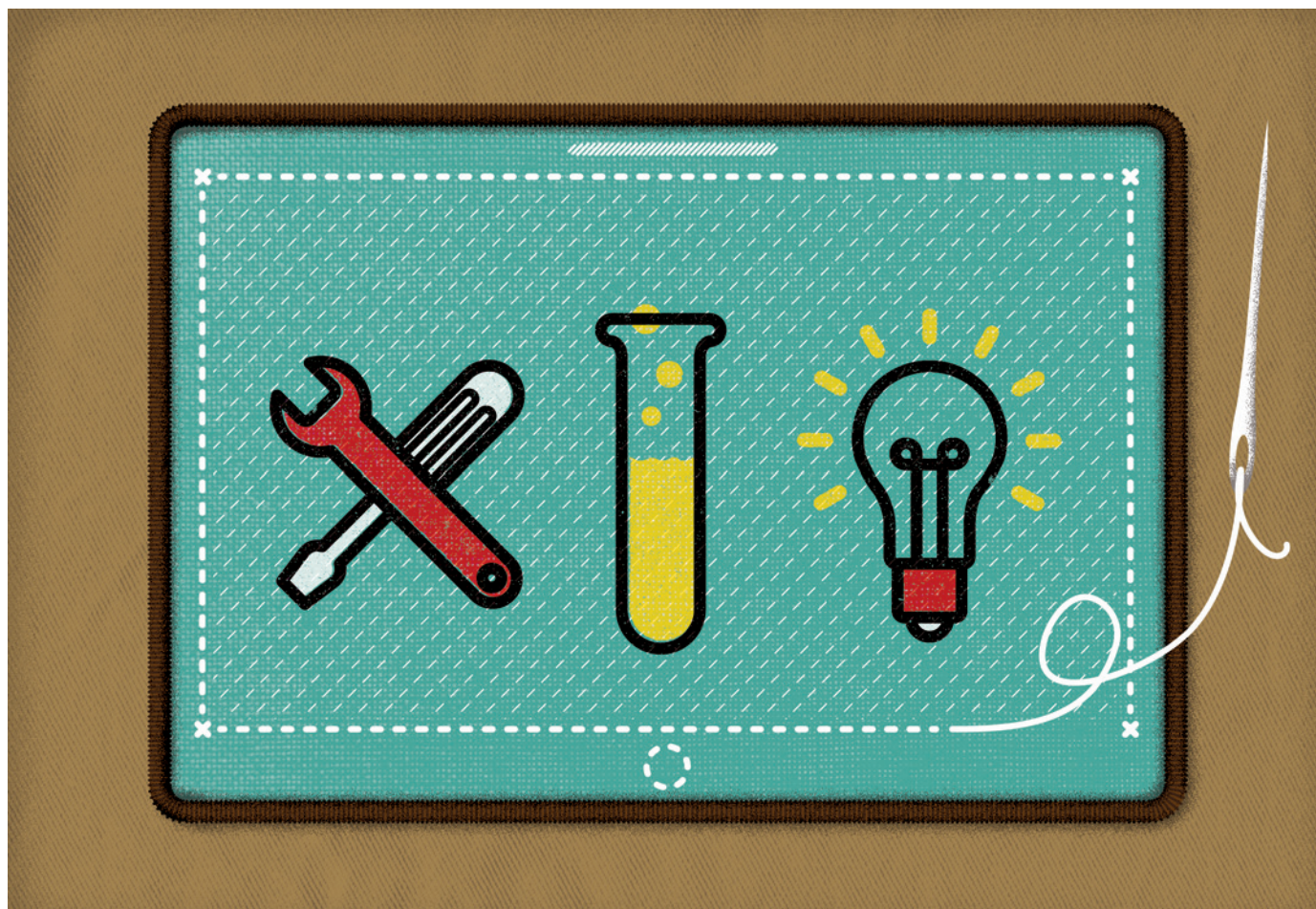
Nature **522**, 197–201 (2015); doi:10.1038/nature14499

In this Letter, the third sentence of the Acknowledgements should have read: “We thank M. Bar-Matthews for the updated version of the Soreq Cave chronology table, R. N. Drysdale and J. Hellstrom for helpful discussions, B. Martrat, P. Jimenez-Amat, R. Zahn, and J. O. Grimalt for the ODP976 records, L. Skinner for the MD01-2444 data, and other colleagues who made data available through the NOAA National Climatic Data Center and/or Pangaea.”. The online versions have been corrected.

TOOLBOX BADGES OF DISTINCTION

A standardized system of digital badges that flag each author's contributions to a research paper aims to enhance collaboration and assign fair credit.

ILLUSTRATION BY THE PROJECT TWINS



BY DALMEET SINGH CHAWLA

An initiative that uses colourful 'digital badges' to denote different contributions to research aims to standardize and simplify the often-fraught business of detailing who did what on a scientific paper.

Two publishers have begun assigning authors any of 14 badges that delineate the parts they played in a study: from a magenta 'Resources' one (for providing study reagents or instruments) to a red one for writing the initial draft. The badges, says Amye Kenall, associate

publisher at BioMed Central in London, could help to minimize the politics of authorship lists, in which supervisors can gain credit for work done by their doctoral students. The project also aims to enhance collaboration by clearly demarcating each contributor's specialities, she says.

On 28 September, the BioMed Central journal *GigaScience* added the badges to two of its published papers. Readers can click to see co-authors listed under multiple badges; the information is also coded in a format that allows computer programs to extract it, which makes it linkable to other online author profiles (such

as the researcher identification system ORCID). Another London-based publisher, Ubiquity Press, is also adding badges to two of its published papers.

"In order for information around contribution to be meaningful and useful, it needs to be standardized," Kenall explains. Many papers include author-contributions sections, but their formats vary, and they can be a vehicle for ambiguity — or insider jokes. In one of the papers with badges, author Keith Bradnam, a bioinformatician at the University of California, Davis, is described in the contributions ►

► section as having “herded goats”.

The concept has been developed by collaborating publishers, research funders and software firms (which have used digital badges for a few years as a visual sign of achievement). Several other publishers have expressed interest in implementing them, Kenall says, and initial feedback from researchers has been positive.

The 14 categories come from a related ‘digital taxonomies’ project, which last year brought together journal editors, funders and researchers to classify authors’ contributions as a set of standard roles (see L. Allen *et al. Nature* **508**, 312–313; 2014).

“We think it’s timely to have a bit more granularity around contributions to scholarly published work,” says Liz Allen, a co-founder of the digital taxonomies project. Accurately determining co-authors’ roles might also help with grant-funding applications, she adds, because applicants could be more explicit about research contributions.

The taxonomy is still in a rough format, but the badges project is not alone in implementing it, Allen adds. *Cell Press*, for instance, now offers researchers the option to use the taxonomy when submitting papers; so far, it has published two papers that do so — although without the badges.

But contributions to scholarly products may be too varied to be captured with a 14-part taxonomy, says Melissa Haendel, who develops systems for querying and classifying biological data at Oregon Health & Science University in Portland. Haendel co-chairs a working group as part of FORCE11, a community-driven initiative that aims to improve scholarly communication technologies and policies. The group is mapping out author roles, in part by using computer programs to search the text of author-contributions sections on papers.

A January workshop in Oxford, UK, listed more than 500 tasks that authors might want to be credited for, she says; examples include developing experimental protocols, taking photographs, developing validated surveys or providing lab reagents.

The badges that a researcher might collect could easily be extended, Kenall notes; extra categories could credit peer reviewers, for example. But for now, BioMed Central is focusing on collecting data about how often people click on the badges, before advancing conversations with funders, publishers and researchers about their practicality, and rolling out the badges to other journals.

“Unlike a CV or author-contributions section, badges provide a method of credit and transparency around contribution fit for purpose for a digital world,” Kenall says. ■

Editor’s note: Dalmeet Singh Chawla worked at BioMed Central until June 2014, but had no involvement with the badges project.

PUBLISHING

The journal that publishes no papers

Mathematics journal ‘overlays’ arXiv preprint server.

BY PHILIP BALL

New journals spring up with overwhelming and almost tiresome frequency these days. But *Discrete Analysis* is different. This journal is online-only — but it will contain no papers. Rather, it will provide links to mathematics papers hosted on the preprint server arXiv. Researchers submit their papers directly from arXiv to the journal, which evaluates them using conventional peer review.

With no charges for contributors or readers, *Discrete Analysis* will avoid the commercial pressures that some feel are distorting the scientific literature, in part by reducing its accessibility, says the journal’s managing editor Timothy Gowers, a mathematician at the University of Cambridge, UK, and a winner of the prestigious Fields Medal.

“Part of the motivation for starting the journal is, of course, to challenge existing models of academic publishing and to contribute in a small way to creating an alternative and much cheaper system,” he explained in a 10 September blogpost announcing the journal. “If you trust authors to do their own typesetting and copy-editing to a satisfactory standard, with the help of suggestions from referees, then the cost of running a mathematics journal can be at least two orders of magnitude lower than the cost incurred by traditional publishers.”

The cost to the journal is only US\$10 per submitted paper, Gowers says; money required to make use of Scholastica, software that was developed at the University of Chicago in Illinois for managing peer review and for setting up journal websites. (The journal also relies on the continued existence of arXiv, whose running costs amount to less than \$10 per paper). A grant from the University of Cambridge will cover the cost of the first 500 or so submissions, after which Gowers hopes to find additional funding or ask researchers for a submission fee.

OVERLAY JOURNALS

The idea of an ‘overlay’ journal that links to papers hosted on a preprint server is not new. There are arXiv overlay journals in maths already, such as *SIGMA* (*Symmetry, Integrability and Geometry: Methods and Applications*) and *Logical Methods in Computer Science*.

But Gowers’ announcement is likely to widen interest in the idea because of his influence in the mathematics community — and outside it.

Three years ago, a blogpost announcing Gowers’ personal boycott of the Dutch publishing giant Elsevier helped to spark the ‘Cost of Knowledge’ movement, which has seen more than 15,000 researchers variously pledging not to publish with, referee for or do editorial work for Elsevier.

And in 2013, Gowers announced his involvement with an initiative called the Episciences project, in which mathematicians decided to launch a series of overlay journals (see *Nature* <http://doi.org/kwg; 2013>). That uses the multi-disciplinary archive HAL, a preprint server that mirrors arXiv and is hosted in Lyons, France. One of its leaders, mathematician Jean-Pierre Demailly of the University of Grenoble in France, admits that progress has been sluggish. “The technical development of the Episciences platform took about a year and a half longer than initially envisioned,” he says. “However, things are now coming along nicely.” The initiative now has five or six staff, Demailly says, and operates three computer-sciences journals and one in maths, which charge nothing to publish.

Episciences would have been a suitable platform to support *Discrete Analysis* too, Gowers says, but he happened to have sufficient funds to use the Scholastica software, and opted for that instead. “I hope that in due course, people will get used to this publication model,” he adds, and that “the main interest in the journal will be the mathematics it contains”. A temporary website on the Scholastica platform will receive submissions before the journal launches early next year.

Gowers says that the model could be extended to other fields. The question, perhaps, is how readily researchers will embrace it. “Apart from being an arXiv overlay journal, our journal is very conventional, which I think is important so that mathematicians won’t feel it is too risky to publish in it,” says Gowers. “But if the model becomes widespread, then I personally would very much like to see more-radical ideas tried out as well” — for example, post-publication review and non-anonymous referees. ■

CORRECTION

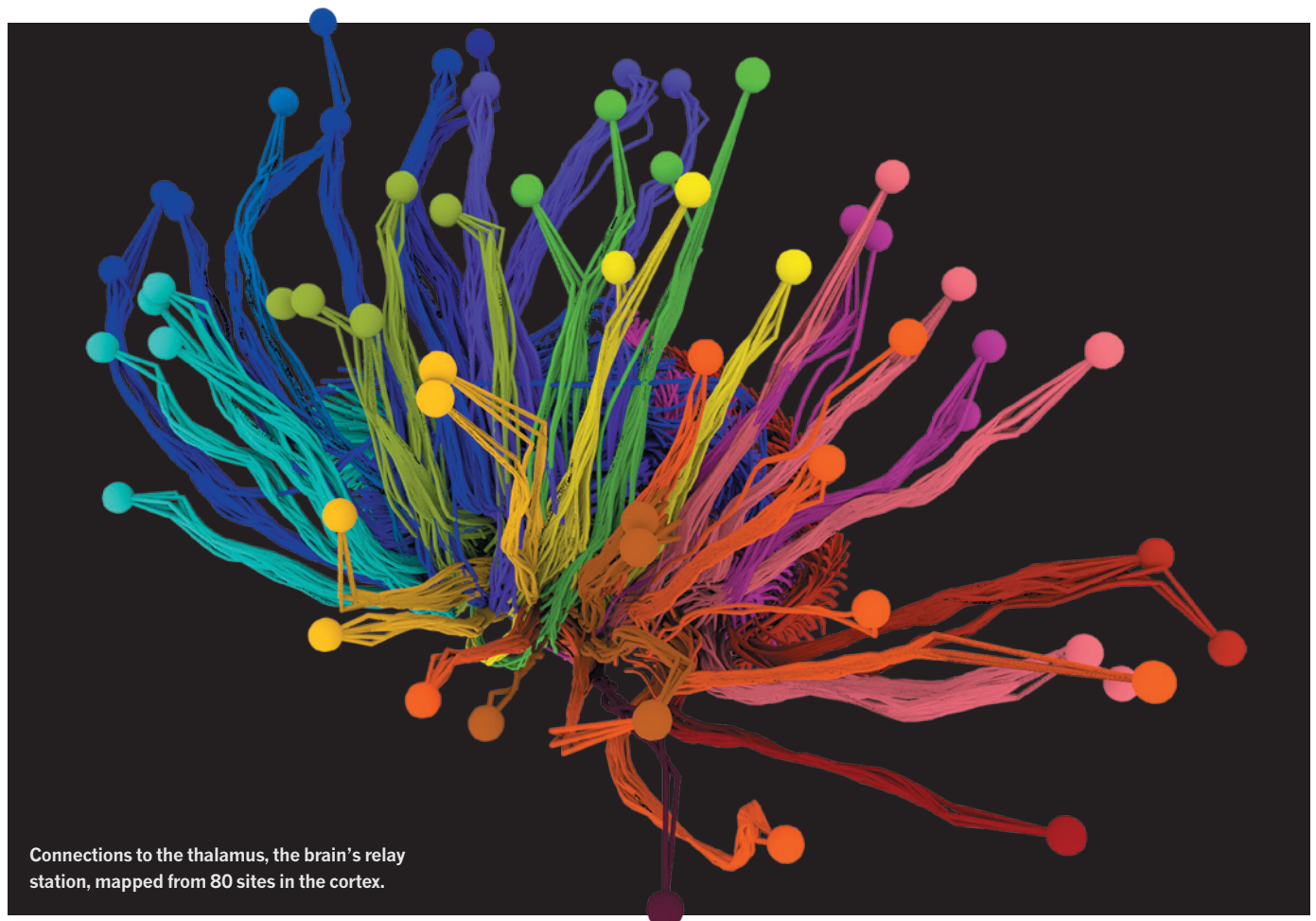
The story ‘See how they run’ (*Nature* **525**, 145–146; 2015) omitted the full name and affiliation for Elizabeth Brainerd, who is at Brown University. It also wrongly stated that Katherine Steele worked on cystic fibrosis instead of cerebral palsy.

TECHNOLOGY FEATURE

CONNECTOMES MAKE THE MAP

Working at a variety of scales and with disparate organisms and technologies, researchers are mapping how parts of the brain connect.

ALLEN INSTITUTE FOR BRAIN SCIENCE



Connections to the thalamus, the brain's relay station, mapped from 80 sites in the cortex.

BY AMBER DANCE

A newborn baby, well fed and sleepy, is swaddled in a blanket and lying on what looks like a tea tray with a helmet attached to one end. Once the infant falls asleep, researchers pull special tabs on the blanket to ease the baby into the helmet. It is a customized receiver coil used for magnetic resonance imaging (MRI), a common method for visualizing brains in living people. The researchers slide the baby-holding contraption along a special trolley into the MRI tube and start collecting images.

From about 1,000 such scans, and another 500 of developing fetuses, UK scientists in the Developing Human Connectome Project plan to map how regions of the brain communicate with each other during development. They then hope to work out why preterm babies are at risk for conditions such as autism spectrum disorder or attention deficit hyperactivity disorder, and perhaps to do similar scans to check whether methods to prevent such disorders are working.

The project is one of many to unravel the 'connectome', the links between the brain's hundreds of areas and millions of neurons.

"The days of just looking at one part of the brain are waning," says Arthur Toga, director of the Laboratory of Neuro Imaging at the University of Southern California (USC) in Los Angeles. He and other scientists are already starting to compare the connections in healthy brains with those of people who have connectopathies, diseases caused by aberrant connections, such as schizophrenia, or disrupted connections, like Alzheimer's disease.

The subjects studied by connectome researchers range from living people to the preserved brains of tiny animals such as worms ►

► and flies. The investigative technologies range from MRI scanners to light microscopes and electron microscopes. Irrespective of the specifics, scientists — with the aid of computers — painstakingly chart connections to build an atlas. The map-makers hope that revealing the connectome's structure will help neuroscientists to navigate as they work out how different parts of the brain function together.

Like traditional cartography, brain mapping is a matter of scale (see 'Maps across magnitudes'). Researchers such as Toga who study the brains of living people are limited to a global view. "It's basically a fly-over at 39,000 feet," Toga says. This approach, called macroscale by some, shows how bundles of axon fibres connect large regions together. With millimetre resolution, it is like a country map that marks major highways. Scientists studying animal brains slice by slice get more detail. At this mesoscale, researchers see how smaller regions of the brain communicate along single axons at micrometre or submicrometre resolution. It is like adding in the lanes of highways and local streets. Finally, microscale images reveal individual neurons and synapses at resolutions of a few nanometres — akin to a map that shows even footpaths and stepping stones.

FLY-OVER

In a major effort to visualize the brain's superhighways, 100 researchers across 10 institutions are close to wrapping up the 5-year, US\$30-million Human Connectome Project (HCP), funded by the US National Institutes of Health¹. By early 2016, they expect to complete MRI scans on 1,200 healthy young adults. They recruited twins — both identical and fraternal — and their non-twin siblings to investigate how patterns in brain connectivity might be inherited; they also collected data such as IQ scores and smoking habits to look for correlations with the connectome. By the end of the project, they will have amassed a petabyte's (10¹⁵) worth of pictures.

HCP researchers image the basic structure of the brain and bundles of axon fibres. They measure blood oxygen levels across the brain as an indicator of activity, looking for areas that fire as people perform tasks or just zone out. Brain areas that are active at the same time are likely to work together.

To get the most information out of each subject, HCP collaborators worked with Siemens Healthcare in Erlangen, Germany, to soup up a standard MRI scanner. It generates a 3-tesla magnetic field — comparable to that in standard machines — but can control the field more precisely. MRI scanners use gradients of magnetic fields to aim at parts of the brain, and the stronger gradient of the HCP machine offers faster imaging and better resolution. That creates more-detailed images of axon bundles. A version of their machine is now available commercially, known as the MAGNETOM Prisma.

Many standard MRI machines collect

images through the brain one slice at a time, but others, including the HCP one, collect eight cross-sectional images of the brain at once, helping researchers see which brain regions are working at the same time. HCP collaborators have also scanned some subjects with a 7-tesla machine, getting even higher-quality data.

In a separate arm of the HCP project, Toga and another group of collaborators are pushing MRI technology in another way. They are improving how machines visualize axon bundles by making use of the restricted movement of water molecules within them. Such diffusion imaging can typically detect water moving in no more than 64 directions. With the HCP's diffusion spectrum imaging software, MRI machines detect hundreds of directions, and thus reveal smaller axon bundles than regular MRI can.

But high-quality images are not enough. To compare images between subjects, researchers use academic-written software such as FSL and FreeSurfer to stretch and squeeze each brain image into a standard shape. The programs must also track how each image was warped, because therein lie key data on what differentiates one human brain from another.

Wide-scale comparisons are planned. The 1,200 young adults, aged 22–35, who were scanned for the HCP are just the beginning. The NIH will now fund projects that look at children and older adults. Combining those connectomes with results from the Developing Human Connectome Project, scientists will have brain maps of the whole human lifespan. The NIH also plans to sponsor work focusing on people with particular diseases or genetic profiles.

Despite the advances, David Van Essen, a neurobiologist at Washington University in St. Louis, Missouri, and co-leader of the HCP, cautions that MRI images can only approximate the wiring of the brain. Scientists working at the mesoscale level get more detail by using light microscopes to look at brain slices.

At this scale, scientists work to pick out groups of neurons and their outgoing axons. Mesoconnectome cartographers therefore inject tracers to label a specific brain region and its conversational partners. Most of the work is in mice, but some researchers are getting started with marmosets, primates the size of kittens.

At the Allen Institute for Brain Science in Seattle, Washington, Hongkui Zeng and her colleagues put together a mesoconnectome by injecting the brains of living mice with viruses carrying the gene for the glowing marker green fluorescence protein (GFP). The neurons at each injection site accumulate GFP along their axons, and so point to the other neurons that they communicate with.

To image the brain, microscopists typically slice it as thinly as possible, but that can mangle

the tissue and reduce the quality of the image. Zeng therefore uses a technique called serial two-photon tomography² in a system called the TissueCyte 1000 that she helped microscope company TissueVision in Cambridge, Massachusetts, to design. In conventional microscopy, GFP requires only one photon to fluoresce, but Zeng's set-up requires the marker to take a double hit. Any wayward photons that flow above or below her plane of focus make no difference to the image, because it is unlikely that two off-target photons will hit the same GFP molecule.

The next trick is that sections are scanned before they are sliced. Researchers embed the mouse brain in a stabilizing matrix of agarose, then image just below the top surface. A cutter integrated into the microscope then shaves off a 100-micrometre-thick section, and the microscope images beneath the new top surface, repeatedly, all through the brain. "It's never damaged before we do the imaging," says Tim Ragan, president of TissueVision. The system can scan a whole mouse brain overnight, Zeng says.

But imaging is only part of the battle. The microscope yields 2D sections of 3D axons that are tangled together. Scientists — or rather, their computer algorithms — must examine each section, tracking the axons through more than 100 images. That can take more time than the scanning.

USC neuroscientist Houri Hintiryan, who is working to generate a mouse mesoconnectome using multiple coloured tracers³, says that the gold-standard tool for this analysis is the human eye. She spends a lot of time lining up structures in sequential images. "That is very exhausting," she says. "However, that is probably the most reliable way to do it up to this point."

Eventually, however, the process will need to be automated, says neuroscientist Partha Mitra of Cold Spring Harbor Laboratory in New York. "One has to build a virtual neuroanatomist," he says. "I want a machine to look at the slides and make sense of them." He and other researchers are working toward this goal. At the Allen Institute, Zeng's team already uses in-house software that quantifies the GFP signal in blocks measuring 25 micrometres per side, about the size of a single cortical neuron, rather than tracing individual cells directly. The team has already processed images from more than 2,100 mice, and made the data available online. It expects to add connection information from hundreds more mice over the next couple of years.

ZOOMING IN

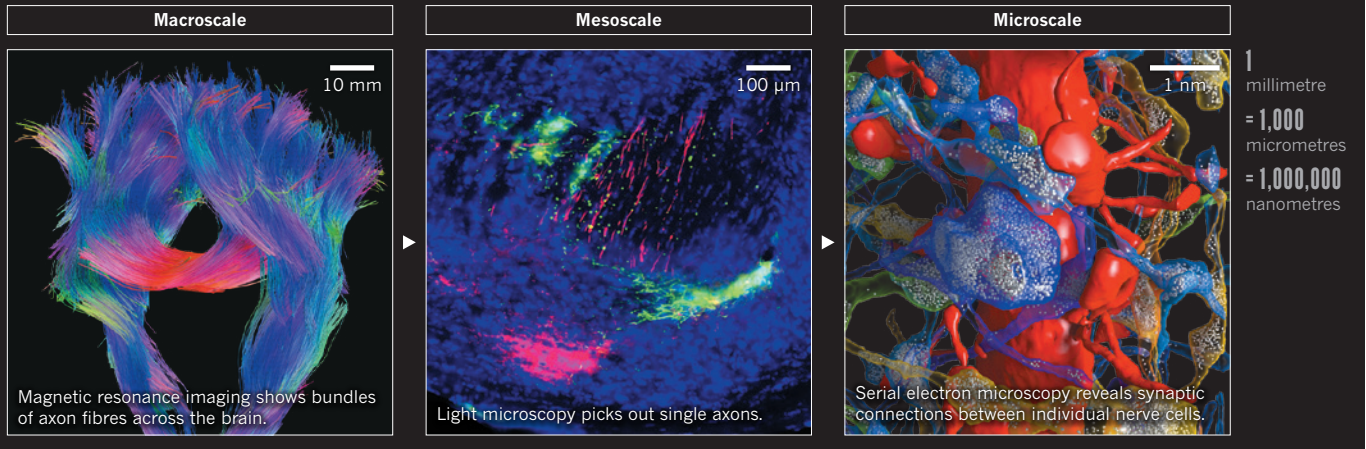
Even the mesoscale connectome offers only part of the brain's story. Microscale neural map-makers want to see connections between neurons — individual synapses where an outstretched axon meets a spiny dendrite. Every neuron talks to thousands of others, so each might have thousands of synapses.

And scientists still want to look at wide sections of the brain. "This is an attempt to

"One has to build a virtual neuroanatomist. I want a machine to look at the slides and make sense of them."

MAPS ACROSS MAGNITUDES

Some connectome researchers track large-scale connections in living brains; others drill down into finer details using thinly sliced brain tissue.



step back, but keep the detail,” says Moritz Helmstaedter, director of the Max Planck Institute for Brain Research in Frankfurt, Germany. “This is why it’s an enormous endeavour.”

For this, researchers rely on electron microscopy. In the Fly EM project at the Howard Hughes Medical Institute’s Janelia Research Campus in Ashburn, Virginia, collaborators use focused ion beam scanning electron microscopy (FIB-SEM) for a serial approach analogous to what Zeng does with light microscopy. They scan the top of a fruit-fly brain, then use the ion beam to sandblast just 8 nanometres off the top before scanning again, then repeat all through the brain for a total of about 500,000 slices.

It takes two or three years to section and image just one fly brain with FIB-SEM, although the researchers can reduce that time by splitting the task between a few microscopes. Fast machines are essential for larger mouse brains, and so Carl Zeiss Microscopy collaborated with connectome scientists to develop the MultiSEM 505 microscope. The device uses not one electron beam, but 61 or even 91, so it can do the work of dozens of electron microscopes at once. Imaging one square millimetre of tissue, in a single plane, takes just eight minutes, says Stephan Nickell, a product manager at Zeiss in Oberkochen, Germany. By tiling images together, users can get a picture representing a slice of brain that is several millimetres, and sometimes even centimetres, across, but can still zoom in for nanometre-scale details⁴.

Again, the hard part is the data processing, and humans still do it best, says neurobiologist Jeff Lichtman of Harvard University in Cambridge. He and his colleagues are working on an algorithm to take over. “It’s about 95% accurate, which is terrible,” he says; he thinks that they can improve on that. Janelia scientists do not fully trust the computer yet, either; they let it make the first pass at identifying cells and synapses, but then use human proofreaders.

Others crowdsource the challenge. For example, Helmstaedter developed a game, called Brainflight, in which players ‘fly’ through the brain’s nerves and software captures those movements to define the borders of the axons. “Even lay people can do it within minutes,” he says.

Helmstaedter and Winfried Denk — director of the Max Planck Institute for Neurobiology in Martinsried — have published the largest microconnectome reported so far: a cube of mouse retina measuring 100 micrometres to a side and encompassing about 1,000 neurons and 250,000 synapses⁵. That was about two-millionths of the mouse brain. Helmstaedter’s next goal is a cubic millimetre of cortex, which is roughly 1,000 times bigger. Denk’s ambition is a full mouse brain.

CONNECTOME IN ACTION

The number and extent of connectomes that are ready for mining will grow quickly over the next decade. Meanwhile, scientists are making headway with the bits and pieces. Thousands have accessed the partial HCP data set, Van Essen says, and Zeng says that thousands visit the Allen connectome database every month.

Neuroscientist Ian Meinertzhagen of Dalhousie University in Halifax, Canada, offers a straightforward example of how connectomics contributed to his work with the *Drosophila* vision system. Fruit flies are attracted to ultraviolet light, and certain photoreceptor cells are known to detect this wavelength. Armed with his electron-microscopy maps, Meinertzhagen predicted that certain neurons in the optic lobe would receive input from those photoreceptors. Sure enough, when his collaborators deactivated those connections, the flies no longer preferred ultraviolet light⁶.

These connectomes will provide fundamental information for many neuroscientists, says Mark Mattson, chief of the Laboratory of Neurosciences at the US National Institute on Aging in Baltimore, Maryland. “It’s important to

know what neurons connect with other neurons in the brain; it’s important to know how much variability there is between individuals.”

But there is still debate about what information is needed, and the level of detail that will be most useful. Tony Movshon of New York University holds that the mesoconnectome hits the sweet spot to understand neural circuits — the level of brain function that neuroscientists most want to understand. For example, scientists interested in how the brain processes sound or touch could follow the mesoconnectome pathways to identify possible members of the relevant circuits. The microconnectome, to his mind, provides too much detail to ask those kinds of questions. At that level, scientists are “doomed to be lost in the forest by looking at all the individual branches”, he says. And the macroconnectome fails to pick up many connections, so scientists will miss important components of the circuit.

But others say that all scales are essential to the next phase of neuroscience, even if it is still too early to predict precisely how. Advances such as the light microscope, and the electron microscope after it, revealed a cellular universe unimaginable by those lacking such equipment, Lichtman points out. The connectome will do the same, he says, even at the microscale. “For that reason alone, looking at brains at this level is likely to be interesting.” Eventually, such information will be a resource that scientists depend on, predicts Denk. “It’s like the genome. It’ll be something that nobody will want to do without.” ■

Amber Dance is a science writer in Los Angeles, California.

1. Van Essen, D. C. *et al. Neuroimage* **80**, 62–79 (2013).
2. Ragan, T. *et al. Nature Meth.* **9**, 255–258 (2012).
3. Oh, S. W. *Nature* **508**, 207–214 (2014).
4. Eberle, A. L. *et al. Micros. Today* **23**, 12–19 (2015).
5. Helmstaedter, M. *et al. Nature* **500**, 168–174 (2013).
6. Gao, S. *et al. Neuron* **60**, 328–342 (2008).

CAREERS

TRANSITIONS From building houses to building molecules **p.153**

FUTURE PLANS Three steps to prepare for the next five years go.nature.com/bpd1rc

NATUREJOBS For the latest career listings and advice www.naturejobs.com



DRA_SCHWARTZ/GETTY

GENETICS

Fluent in DNA

As genomics migrates to the clinic, job options are emerging for genetic counsellors to explain the meaning in mutations.

BY MICHAEL EISENSTEIN

Jehannine Austin spent years at the laboratory bench developing techniques for detecting genetic mutations, and was well into her PhD programme before dissatisfaction began to set in. “I was working on the genetics of schizophrenia and bipolar disorder, but I didn’t have the communication skills or knowledge to explain how what I was doing was relevant at a personal level,” she says. She wanted to help patients and their families to understand the potential medical implications of genetic changes. And so, after earning her

degree, she enrolled in a master’s programme in genetic counselling at the University of British Columbia in Vancouver, Canada.

Clinicians routinely recommend genetic tests as part of diagnosing disease and predicting disease risk, but many people find it hard to understand exactly what it means when a result indicates greater risk or the potential to pass on diseases to their children. Genetic counsellors help individuals to see how their genomic variation could affect their own health and that of their family members. A counsellor’s role is to explain what testing may or may not reveal, and how to prepare for the

medical and psychological implications of those results. That includes helping people to respond proactively to findings — for example, genetic counsellors often encourage early cancer screening for women who have a *BRCA1* mutation that strongly elevates their risk of breast cancer.

The ranks of genetic counsellors have swelled in recent years. Throughout the United States and Canada, there are more than 3,500 licensed practitioners, according to the American Board of Genetic Counseling (ABGC), the certifying organization for these two countries. And at least another 300 work in the United Kingdom. But there are more jobs than eligible candidates: a 2012 assessment from the US Bureau of Labor Statistics determined that prospects in the field are growing much more rapidly than the average for other occupations.

FROM LAB TO CLINIC

A genetic counsellor’s job straddles the fields of biology and psychology, and the position draws many scientists who are looking beyond the bench. “When people apply to genetic-counselling programmes, their letters typically say, ‘I love genetics, but the lab isn’t for me — I want to be at the people end of things,’” says Austin, now a genetic-counselling faculty member at her alma mater and the incoming president of the US National Society of Genetic Counselors.

Those who wish to pursue the career need a master’s degree in genetic counselling and a certification or licence to practise. Professional societies administer such certification in Australia, North America, Japan, South Africa and the United Kingdom; in mainland Europe, individual programmes handle licensing. Worldwide, graduate programmes emphasize a similar foundation of genetic-literacy and counselling skills paired with hands-on experience in the clinic. Training focuses on the situations that most commonly require genetic counselling, including hereditary cancer risk, prenatal diagnosis of birth defects and early-onset or developmental disorders in children.

The first genetic-counselling programme was established in 1969 at Sarah Lawrence College in Bronxville, New York. For decades, the job centred mainly on understanding patterns of inheritance, and attempting to diagnose genetic maladies on the basis of physiological or biochemical indicators. But the field of clinical genetics shifted dramatically in the 1990s as technological advances began to ►

► enable physicians to interpret the medical consequences of specific sequence variations, such as the link between mutations in two *BRCA* genes and the risk of breast and ovarian cancer. The field advanced further in the 2000s with an explosion in genomics — next-generation sequencing platforms for fast, low-cost genetic analysis. This greatly accelerated the discovery of gene–disease associations, and gave clinical geneticists a much deeper well of knowledge from which to draw.

AT EASE WITH UNCERTAINTY

Today, most genetic counselling is clinical, focusing on helping patients to understand the consequences of mutations either in individual genes or in panels of specific genes that have ties to conditions such as cancer or neurological disease. But a growing number of academic centres, hospitals and private companies are exploring genome-scale sequencing and the extent to which information about thousands of genes can guide clinical diagnosis of inherited diseases and cancer. Some countries are pursuing national-scale medical-genomics programmes, such as the ‘100,000 Genomes’ initiative in the United Kingdom, which formally began recruitment early this year (see go.nature.com/ri9rn5).

That translates into an unprecedented demand for genetic counsellors who can tackle the surge in the number of people who are choosing to have their genomes analysed. Professionals must grapple with the challenge of results from thousands of genes simultaneously, rather than a carefully selected — and well-understood — handful.

The shift to large-scale sequencing does not necessarily rewrite the counsellor’s job description, but it does add complexity. One major issue is uncertainty — genome-scale analysis still routinely fails to definitively identify a causative mutation, and may instead return ‘variants of unknown significance’. These reside in genes that are thought to be clinically important, but they have not been clearly

demonstrated to interfere with gene function.

Such mutations are relatively common, says Kelly Hagman, managing director of clinical genomics at Ambry Genetics, a genetic-testing company in Aliso Viejo, California. She estimates that 10% of Ambry’s ‘exome tests’ (which sequence all the protein-coding genes in a given genome) return these undetermined variants. Although the emotional effects of this uncertainty has been a major source of concern, genetic counsellors generally find their patients to be resilient, mainly because genomic analysis is often just the latest stop on a years-long diagnostic odyssey.

More challenging is the potential for ‘incidental findings’ that might be uncovered in a person’s genome while searching for a specific clinical result, such as assessing cancer risk. To try to pre-empt undesired, disturbing disclosures, patients are now asked before undergoing sequencing whether they would want to receive such results.

Still, the array of results and decisions can make the post-sequencing meeting harder for a counsellor to navigate. “Previously, you knew the condition you were testing for when the family arrived, and you could counsel them on it,” says Sarah Scollon, a genetic counsellor involved with a large-scale clinical exome-sequencing programme at Baylor College of Medicine in Houston, Texas. “Now you’re throwing a wider net and can’t do that kind of specific counselling up front; that now comes more at the back end.”

TO THE LAB AND BEYOND

On the plus side, the exploding interest in clinical genomics has created a wealth of opportunities for genetic counsellors. The boom in exome- and genome-research programmes, and the rapid growth of commercial sequencing providers such as Ambry, have shifted the field to applied research: now, genetic counselling beckons even those who most enjoy the bench. “When I first started around 2004, less than 10% of counsellors worked in labs, and now it’s well over 20%,” says Hagman.

At Geisinger Health System, a large health-care provider in Danville, Pennsylvania, the counsellor workforce has grown by 75% since 2006, according to Janet Williams, the company’s director of research genetic counsellors. And Hagman reports that Ambry employs 70 or 80 genetic counsellors, and is “always hiring”.

Many of these counsellors have little direct contact with patients; instead, they write educational materials for physicians and patients, explain the technology to clients, consult on test selection and publish research. These

counsellors are also working on cutting-edge capabilities, adapting to rapidly evolving genetic knowledge. Across Ambry’s exome tests, roughly one-quarter of results flagged in diagnostic sequencing are mutations that were discovered to be clinically relevant only in the past two years, says Hagman.

Public-education and policy specialists who have genomics expertise are also in demand, says Barbara Biesecker, who directs a joint genetic-counselling programme between Johns Hopkins University (JHU) in Baltimore, Maryland, and the US National Human Genome Research Institute (NHGRI). She notes that the programme helps students to explore careers in policy and industry.

There has also been discussion about creating advanced-degree programmes for genetic counsellors who are interested in further specialization. A task force formed by the ABGC in 2011 found that counsellors want more opportunities for academic growth. “There is a lot of interest in the idea of genetic counsellors gaining higher-level, research-based training,” says Austin. In 2013, the ABGC followed up on this to form the Committee of Advanced Training for Certified Genetic Counselors, aiming to assess the skills and education that counsellors may require for career tracks outside the clinic.

Some worry that the rate of training is failing to keep pace with demand. The 34 accredited genetic-counselling master’s degree programmes in the United States and Canada attract many more applicants than they can accommodate. The prestigious 2.5-year JHU–NHGRI programme, established in 1996, receives 80 to 100 applicants a year, yet accepts only 4 or 5 students. Several more programmes are under development, but funding is tight — multiple programmes have folded in past years because of budgetary constraints.

The United Kingdom offers only two training programmes, and in most other countries, the field is just now emerging as a career option. “They’ve either had no counsellors or they’ve had medics or nurses doing genetic counselling,” says Anna Middleton, a genetic counsellor at the Wellcome Trust Sanger Institute in Cambridge, UK.

Nonetheless, the profession is expected to become more deeply embedded in a wide range of medical and scientific settings. A July paper by a team from the American Society of Human Genetics (J. R. Botkin *et al.* *Am. J. Hum. Genet.* **97**, 6–21; 2015) cited the low number of counsellors as one of the barriers hindering the incorporation of genetic medicine into standard practice. “I think that in medicine, for a long time, we’ve taken the approach that the information itself is all you need, but that’s blatantly wrong,” says Austin. “It is exquisitely important how you deliver that information.” ■

Michael Eisenstein is a freelance writer in Philadelphia.



Sarah Scollon advises people about gene testing.

STEPHANIE GUTIERREZ

TURNING POINT

Hosea Nelson

A former construction worker, Hosea Nelson is still building things — but now, as a chemist, he constructs biologically active molecules instead of houses. In July, he launched his own laboratory at the University of California, Los Angeles (UCLA).

PENNY JENNINGS



Can you describe your path into science?

I did construction and odd jobs for years, but over time, I decided I wanted to go to school. The problem was that I had no idea what to do. I took biology courses at a community college and liked them, and eventually, I got a job doing research in a microbial-genomics lab at San Francisco State University in California. I fell in love with research and the pursuit of intellectual ideas. Making a genetic construct gave me the same feeling as building things. But I wanted to understand the glue — the chemical bonds that hold everything together — and got more interested in chemistry and building molecules. I spent four years at a community college, then transferred to the University of California (UC), Berkeley, and knocked out a bachelor of science degree in chemistry in two years.

How did you decide where to go for your graduate degree?

I didn't go right away. I got a job in industry at the Panasonic Energy Solutions Lab in California's Silicon Valley. One of my UC Berkeley instructors stayed in touch, and convinced me that I was talented and could get my costs covered for graduate school. I had never heard of the California Institute of Technology in Pasadena, but I applied there and worked with chemist Brian Stoltz, who is known for making compounds relevant to human health.

Do you feel you charged ahead in Stoltz's lab?

Yes. He and I are similarly intense, but he is

much more regimented, whereas I grew up in San Francisco with hippie parents who did not have a lot of structure. I was worried that we would clash, but we share a common interest in constructing molecules. In his lab, I decided to do a total synthesis of molecules derived from a plant called *Thapsia garganica*, which has a rich medicinal history that even Hippocrates wrote about.

To what do you attribute your success?

I take a lot of initiative and have a fearless approach, which I learned from the industrial lab. The biggest roadblock to productivity is indecision. If you have an idea, no matter how crazy you might think it is, you have to go with it to get anything done. With the *T. garganica* project, I spent 3 months trying to find a way to make a compound using a conventional route, which would have been a 20-step process. I challenged myself to do it in six or seven steps, and my short route ended up working.

Did you ever feel out of place in academia?

Absolutely. I still do at times. Culturally, socio-economically and educationally, I came from a different place from most high-level academics, who often follow a similar path from a similar socio-economic background. I was a high-school dropout and poor. I have a high level of appreciation for being able to put food on the table and for the fact that I love my job, but I'm not sure that appreciation is universal.

Has everything clicked into place easily for you?

No. Soon after I started a postdoc to work on catalysis at UC Berkeley with a funded proposal in hand, my idea got scooped. I spent three months trying every wild idea that could be related to my proposal so that I wouldn't lose my funding, and one worked well. A couple of others looked promising. It was validation. It taught me that I'm good at coming up with project ideas. I ended up forming my own research group within my adviser's group, and convinced the team to work on how to use ion pairing — oppositely charged ions that behave as a single particle — to speed up chemical reactions. I published a lot, and spent time working with undergraduates and graduate students.

Was your adviser OK with that?

Most successful researchers do not argue with

good results. I convinced him to let me do this after I gave him results that he deemed worthy of investigating.

Did you learn about lab management?

Yes. I am more intense than the average postdoc. I think about chemistry in the shower. One thing I found out quickly is that not every researcher is that obsessed. I realized I needed to temper my enthusiasm, so that I would not come off as a tyrant.

How was the job search?

Gruelling. I am not used to travelling to three different universities in one week and giving six lectures. It is especially hard at places such as the Massachusetts Institute of Technology in Cambridge, where you are presenting ideas in front of the best people in the field. I did that for two months and ended up with three offers. It was a complicated decision, but I went with my gut and chose the position at UCLA.

Is there anything you learned as a construction worker that you use in your career now?

One thing that governs how I do organic synthesis is 'measure twice, cut once' to avoid wasting wood. I try to measure twice in a chemical sense before we do anything that is not fixable.

What has been the hardest part of getting your own lab off the ground?

The amount of time it takes to provide adequate supervision for my high level of safety expectations. I have seven people so far — four graduate students, two undergraduates and a postdoc. People make mistakes. Providing the heavy levels of training up front on how to do things safely is worth it, but it takes a lot of time.

What are your goals?

On the teaching side, I'm excited to work with a colleague who is an award-winning organic chemist, Neil Garg, who instructs one of the most highly sought-after classes at UCLA. He is a real innovator, but the class's weak spot is the organic-chemistry lab section, so we are going to teach that together. Once I have established myself and secured tenure, I want to spend time making our field more inclusive. For research, we are starting with five unrelated projects. We're taking the pasta-cooking approach — whatever sticks is what we eat. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for length and clarity.

CORRECTION

The Careers Feature 'Plight of the postdoc' (*Nature* **525**, 279–281; 2015) gave the wrong credit for the photo of Anna Kalashnikova. Credit should have gone to Katherine Labbe.

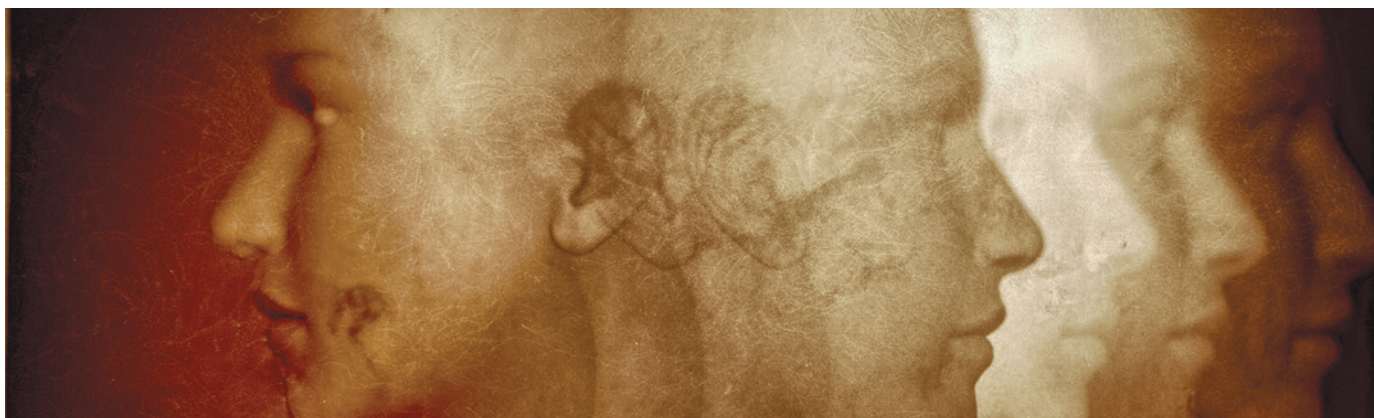


ILLUSTRATION BY JACEY

CORRIDORS

This is where you live.

BY RAHUL KANAKIA

You live in a city that is a planet. Humanity has reached its full maturity. The Galaxy is empty. The Galaxy is full. Your life is built atop a trillion tonnes of steel that your forefathers pulled from the core of planets that are now black and empty.

When you go to sleep, you hear your neighbour arguing with his wife through the wall. He wants her to become a dancer, and she wants to sleep for a thousand centuries. You wish you had someone to argue with.

You live in a city that spans a planet. You go to the gardens on the surface, but you don't stare at the greenery or at the Sun. Instead, you look down into the latticework, at the millions of miles of tubes and wires and structural shapes and spidersteel webbing.

A woman is shrieking and crying, and you wonder why the medical robots haven't yet injected her with a dose of reality, but she says, "No, no, I won't suck down any more of your nasty drink! I've already lost too many years!" and you realize that she is not insane. She is simply bereft.

You live in a city that covers a planet. And within that city you live in a box that has one thousand square feet of surface area. The box has windows, and when you stand next to those windows, you see either: 1) the dark of the inner corridors; 2) the neon gleam of the night districts; or 3) the Sun beaming down through the interstices above.

All three views are the same, of course.

Your mother lives in a chip implanted above your left eyebrow, and your father shot himself into the atmosphere before you were born. You have two clone brothers who look nothing like you, even though everyone says they do. Chalub is the innamorato of a

not-yet-famous composer, and Zohern lives in a stairwell 2,000 miles from here.

The next time you go to the garden, it has acquired a yellow tinge, and you wonder if it's dying. All things die, eventually. That is something your mother told you a long time ago, but she must not have believed her own words, because when you repeated them back to her, she went silent for many years.

Even now, she does not often emerge. You can feel her, basking in the sunlight, thriving on your blood. And her silence is not an awful thing. Some people — their heads sparkle in the sunlight — are forced to wear the chips of a thousand others. They are never free of other minds, other voices. You are not like that. You are almost allowed to be yourself.

A shrill voice pierces the garden, and you look up. The woman is there, carrying on, and you avoid her eyes because you know that if you see each other too many times, then you and she will order a love elixir and choose to drink it with each other, because that is how the city works.

You live on a planet that is covered by a city, and the city has nine trillion inhabitants, and if you repeatedly bump into one of them, then you know that the city has chosen for you.

This is a mystical belief. The city bureaucrats insist that they do no choosing. They insist that the world is cruel and meaningless, and they say the only order is that which arises from a rigidly logical mind.

The woman falls to her knees and shrieks, "A-we! A-we! A-we!"

You are crying. But it is okay, because sometimes these tears are tears of joy. The maybe-crazy woman

kneels opposite you, and the flow of traffic passes between you. A child — she has wisps of wire hanging from her eyebrows — turns around and around and you expect her to fall over, dizzy, but she doesn't. She simply gets faster and faster and faster, until she explodes with light and is gone.

When you were young, you believed that disappearing children went to a better place, but then your mother said no — the children were dead, and death equalled nothingness.

Years later, though, you learned the truth: no one knows where the children go, or why they leave.

You see the girl's father, standing next to the empty space, and her father looks at you. Then he gathers up his daughter's clothing and zips it into a bag, and he walks onwards too.

The woman — the woman who cries and might be crazy — stares at you.

"The city will crumble," you say, but the woman is too far away to hear it.

A-way, A-way, A-way. Some say the world will go on and on, extending into the foreverness, but you don't believe it.

In the meantime, you know that you are lonely. You know that the city is full of lonely people. You know that someday you will meet another lonely soul. And you know that you and she (or you and he) will drink the devil's drink and bind yourselves together in an intimate union that will go on and on and on for centuries until some sideways and mysterious shift of the ghost-brain turns you back into strangers. ■

Rahul Kanakia is the author of a young-adult novel, *Enter Title Here, that will be out from Disney-Hyperion in August 2016.* He lives in Berkeley, California.

➔ **NATURE.COM**
Follow Futures:
@NatureFutures
go.nature.com/mtoodm